THE CONJUNCTIVE FRACTION OF EMBODIED BENCHMARKS: COMPOSITION CANNOT BE MEASURED WHERE THE WRITE PATH HAS ALREADY DESTROYED IT

Anonymous authors
Paper under double-blind review

ABSTRACT

The Capability Convergence Hypothesis (CCH) holds that capability is purchased by *access structure*: the joint possession of a compressive bounded-state channel and a scalable verbatim-index channel. It proves composition strictly super-additive on a witness family that obstructs each channel separately, and names as its largest open risk a measurement it reports "has not been run by anyone, including us": how often natural workloads instantiate that joint load. We run it for embodied multimodal streams. An arithmetic criterion, computable before any run, first decides whether a benchmark *can* exhibit the effect at all: a bounded score cannot host an interaction whose components already sum past the bound. A census of public video question-answering corpora then puts the conjunctive fraction in the low single digits, so solving that subset perfectly would move the richest benchmark by less than the run-to-run floor at which these benchmarks are compared. We supply the witness the hypothesis implies (735 judge-free items over head-mounted kitchen streams), and in the registered grid every prediction fails and the channel dichotomy reverses. A diagnostic locates the failure upstream of every arm, in the shared single-frame write path, which decides the locator verb only 11–23% of the time, so no read-side composition had anything to compose. Supplying the ledger from annotations instead, composition by address beats a same-cost different-occurrence control by 14–26 points on every backbone and is null on both single-axis families (the corpus is a valid witness and the loss is on the write), yet even then it answers less than half.

1 INTRODUCTION

The Platonic Representation Hypothesis holds that as models scale, heterogeneous networks converge on a shared representation (Huh et al., 2024). The Capability Convergence Hypothesis (CCH) proposes its boundary: under a fixed per-token inference budget, representational convergence does not entail capability convergence, and capability instead converges toward the *access-complete hybrid*, any architecture holding both a compressive $O(1)$-state channel and a scalable verbatim-index channel (Chen et al., 2026). Its proved half rests on three resource walls and one witness: a task family no pure architecture solves at any scale while their composition does, on which capability is strictly super-additive.

CCH also names, three times and in its own words, the measurement that would decide whether any of this matters. Its Alternative Explanation 3 concedes that if witness-type loads have vanishing measure in natural workloads, CCH's capability content lives only in synthetic separator tasks, and states that *"the incidence measurement that would decide this has not been run by anyone, including us."* Its Limitation (iii), *"the largest open risk"*, says the uncertainty "is not whether the separation exists but how often natural workloads instantiate the joint load." CCH cannot run it: it lives in symbolic token streams, where the words *vision*, *video*, *multimodal* and *embodied* do not occur, and its own diagnosis points away from language ("S_5-level composition is not frequent in everyday language"), because text is pre-compressed by a writer who supplies the locator with the question. Perception is not. A companion transported CCH to embodied multimodal perception, proved four

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.

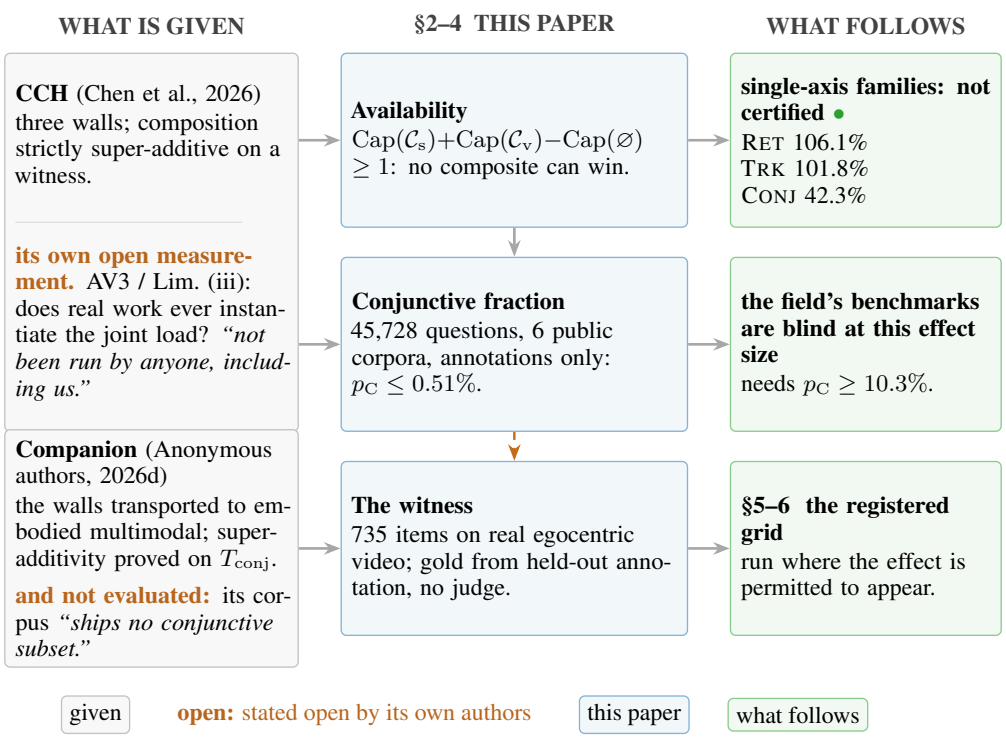


Figure 1: The argument. CCH proves that composition can be strictly super-additive, and names as its own largest risk the one measurement that would say whether the purchase is ever necessary, reporting that nobody has run it. A companion transports the walls to embodied multimodal perception, proves the conjunctive proposition, and then meets a corpus that cannot test it. The gap was never the theory. This paper supplies the three things that close it, and the first is arithmetic rather than experimental: on any query family where one channel alone is near ceiling, strict super-additivity is not merely hard to detect but *unavailable*, so a benchmark can forbid the effect it is being used to look for. The dashed arrow answers the circularity charge the design has to meet: the witness's conjunctive form occurs in corpora nobody here wrote; it up-samples a measured tail rather than inventing a task.

walls in perceptual form and proved strict super-additivity on a conjunctive witness (Anonymous authors, 2026d); it then could not test that proposition, and says so in print, because its licence-free synthetic corpus "ships no conjunctive subset." The gap in this line of work was never the theory. It was a corpus. This paper closes it, and finds that the reason the corpus was missing is itself the result.

The question without the theory. A reader who grants none of the above still faces a question about evaluation. Composite memory, a compressed running summary beside a retrievable archive, is the default design for long-horizon agents, and the standard evidence for the second store is a benchmark delta. Two facts about that practice are checkable without any hypothesis: a bounded score cannot host an interaction whose components already sum past the bound; and whether a given benchmark hosts one is countable from annotations before anything is run. Sections 2 and 3 are those two checks, and nothing in the criterion, the census or the witness depends on CCH being true.

The arithmetic that comes first. Write $\text{Cap}(\cdot)$ for accuracy under a fixed per-decision budget. Strict super-additivity in additive form is

$$\text{Cap}(\mathcal{C}_s{+}\mathcal{C}_v){-}\text{Cap}(\mathcal{C}_v) \;>\; \text{Cap}(\mathcal{C}_s){-}\text{Cap}(\varnothing) \quad\Longleftrightarrow\quad \text{Cap}(\mathcal{C}_s{+}\mathcal{C}_v) \;>\; \text{Cap}(\mathcal{C}_s){+}\text{Cap}(\mathcal{C}_v){-}\text{Cap}(\varnothing), \tag{1}$$

so, since no capability exceeds 1, the inequality is *unavailable* (unsatisfiable by any composite, however good) whenever $\text{Cap}(\mathcal{C}_s) + \text{Cap}(\mathcal{C}_v) - \text{Cap}(\varnothing) \geq 1$. The condition is a property of the *query family*, not of the system: a benchmark can forbid the effect it is used to look for. It is one-sided, because the quantities available before a run are upper bounds: ceilings licence "available", only measured sums licence "forbidden". On the corpus we build, the ceiling sum is 106.1% on episodic

retrieval and 101.8% on state tracking, so neither single-axis family is certified; on the conjunctive family it is 42.3%, which the composite must exceed and which is our sharpest registered prediction.

How much of the field's evidence is single-axis? An instrument that reads annotations only and touches no frame (§3) classifies a query as *conjunctive* when the answer is a verbatim percept at one moment *and* the locator of that moment is itself a reduction over the stream: "what colour pen did the author use when he wrote 'guitar' for the second time." Either half alone is single-axis. Over 45,728 questions from 6 public corpora the conjunctive fraction is at most 0.51%, and VSI-Bench contains none at all. With the detectability arithmetic of §3.1 this says more than "the effect is small": these benchmarks are blind, on a measured lower bound, at the effect size access structure can produce.

And then the witness, on real pixels. Section 4 builds it from EPIC-KITCHENS-100 (head-mounted kitchen video under CC BY-NC 4.0 with no signed agreement, Damen et al., 2022), whose dense start/stop-timestamped narrations let a conjunctive item be generated with mechanically verifiable gold: the locator is the k-th occurrence of an event type, the value is what the person interacted with immediately afterwards, and the annotation is never shown to the model. 735 items over 8 streams; gold is an option letter, so nothing is judged.

Contributions. (1) **The availability condition, per query family** (§2): an arithmetic criterion, computable before any run, for whether a benchmark *can* exhibit strict super-additivity, with the decomposition naming the two factors that decide whether a composite uses the room, converting "hybrid memory does not help" from an empirical puzzle into a statement about what was measured. (2) **The conjunctive fraction of the field's corpora** (§3): the incidence measurement CCH names as unrun, for embodied multimodal streams, with the detectability arithmetic that makes $p_{\mathrm{C}} \leq 0.51\%$ a claim about instruments rather than architectures. (3) **A conjunctive witness on real egocentric video** (§4): generated, gated against degenerate gold, judge-free. (4) **An honest ledger of which bound binds**, including the Shannon wall, *vacuous* here at $\kappa_s = 4.0$–7.4, and the occurrence semantics our own generator got wrong, corrected and re-run. (5) **Witness validity** (§6.2): with the ledger supplied from the annotations and the same budgeted read, the composed arm beats an equal-cost wrong-address control by $+14.3$–$+26.3$ points on all 3 backbones (Holm $p < 10^{-5}$), and is null on both single-axis families. We do *not* test CCH's convergence conjecture: its classes are architecture families, ours are *context-assembly* channels on frozen models.

2 When a Benchmark Can Exhibit Super-Additivity, and When It Forbids It

Setup. A stream $x_{1:N}$ of N keyframes is presented once; a query q arrives after it and must be answered inside a fixed per-decision budget B of context tokens. Two channels may spend it (Chen et al., 2026; Anonymous authors, 2026d): the **compressive channel** $\mathcal{C}_{\mathrm{s}}$, a bounded running state of L_s tokens updated once per keyframe and never re-reading a frame, carrying M bits regardless of N; and the **index channel** $\mathcal{C}_{\mathrm{v}}$, a lossless off-context store of all N keyframes from which a query-conditioned retrieval lifts $w \ll N$ frames into the budget. Both are ways of *filling a context window* on a frozen model, not architectures. $\mathrm{Cap}_B(\cdot)$ is accuracy at B, $\mathrm{Cap}_B(\varnothing)$ the blind agent (nominally $1/K$ on a K-way item); on a query family T we abbreviate $b = \mathrm{Cap}_B(\varnothing|T)$, $s = \mathrm{Cap}_B(\mathcal{C}_{\mathrm{s}}|T)$, $r = \mathrm{Cap}_B(\mathcal{C}_{\mathrm{v}}|T)$, $\sigma(T) = s + r - b$, with $\bar{\sigma}(T)$ the *ceiling sum* when s, r are replaced by upper bounds and b by $1/K$.

Definition 1 (Conjunctive query family)**.** A query q is *conjunctive* if $a(q) = V(x_{L(x_{1:N})})$ where (i) the **locator** L is a reduction over the whole stream (a count, an ordinal, a superlative), so its value is not a function of any $o(N)$ content-selected sample; and (ii) the **value** $V(x_t)$ is a percept at a single moment, not recoverable from a compressive summary that did not know t when it passed t. A query failing (i) is *index-shaped*; one failing (ii) is *state-shaped*. Only a query satisfying both obstructs each channel separately.

Definition 2 (Ledger composite, and its two events)**.** A *ledger composite* $C = (\omega, \rho)$ holds both channels and factors through a ledger: the *write map* ω produces $\ell = \omega(x_{1:N})$ (the state of $\mathcal{C}_{\mathrm{s}}$ plus the entries indexing the frames of $\mathcal{C}_{\mathrm{v}}$), one entry per keyframe, without lookahead; the *read loop* ρ receives (ℓ, q), may retrieve through $\mathcal{C}_{\mathrm{v}}$ under B, and returns an option. On a conjunctive item write $W = \{\lambda(\ell, q) = L(x_{1:N})\}$ and $R = \{\rho(\ell, q) = a\}$, λ being the composite's own locator resolver.

Both events are decidable per item from a cached ledger and the corpus annotations, which is what makes the propositions below measurements.

Proposition 1 (Availability)**.** *On a query family T, (1) admits a solution only if $\sigma(T) < 1$. When $\sigma(T)$ reaches 1 no composite is strictly super-additive on T, whatever its design, because $\mathrm{Cap}_B(\mathcal{C}_\mathrm{s}{+}\mathcal{C}_\mathrm{v}) \leq 1$.*

The proposition is immediate, and that is the point: a hybrid's failure to be super-additive on a benchmark may carry no information about the hybrid. The companion's uniform witness mixture had $\sigma = 1.28$, unsatisfiable *as arithmetic*; what nobody had done is compute the quantity per family.

Proposition 2 (Availability is necessary and never sufficient)**.** *Let $0 \leq \sigma(T) < 1$ and let C be a ledger composite with $p = \Pr[W]$, $c_1 = \Pr[R|W]$, $c_0 = \Pr[R|\neg W]$. Then (i)* $\mathrm{Cap}_B(C|T) = pc_1 + (1-p)c_0$*; (ii)* write-side identity failure*: if $c_0 \leq \sigma(T)$ and $p \leq (\sigma(T) - c_0)/(1 - c_0)$, C is not strictly super-additive on T whatever the read loop does on the items it resolves; (iii)* read-side percept failure*: if $c_1, c_0 \leq \sigma(T)$, the same holds whatever fraction of locators the write resolves; (iv) if $c_0 \leq \sigma(T)$, strict super-additivity implies $c_1 > \sigma(T)$ and $p > (\sigma(T) - c_0)/(c_1 - c_0)$; (v)* not sufficient*: for every such T and C there is a C' with the same write map and $\mathrm{Cap}_B(C'|T) \leq \sigma(T)$, so $\sigma(T) < 1$ decides nothing about any composite.*

Availability is a statement about a *family*; whether a composite clears the bar has exactly two factors by (i), write-side identity and read-side percept, neither bounded below by Proposition 1. §6.1 measures the first.

Assumption 1 (The score is uninformative about the ordinal)**.** Let an event type occur at $t_1 < \cdots < t_m$. A retrieval ranking by a score $s(q, x_t)$ that depends on the query and on frame content ranks the j-th occurrence first among the m no more often than chance, $1/m$.

Proposition 3 (An ordinal locator is not retrievable)**.** *Under Assumption 1, on a conjunctive item whose locator is the j-th of m occurrences and whose K options include the values at all m occurrences as distinct entries, any agent holding $\mathcal{C}_\mathrm{v}$ alone has $\mathrm{Cap}_B(\mathcal{C}_\mathrm{v}) \leq 1/m$.*

This is the round wall of Anonymous authors (2026d) specialised to a form whose ceiling is a number rather than an impossibility. On an index-shaped family $m = 1$ and the ceiling is 1: the proposition that bars $\mathcal{C}_\mathrm{v}$ on the conjunctive family *permits* it everywhere else, which is the dissociation and not a weakness. On the compressive side the natural bound is the Shannon wall ($\kappa_s = (M/(Nd) + 1)/\log_2 K$; Anonymous authors, 2026d), and on this corpus it says nothing: a kitchen annotation carries one queryable attribute per keyframe, so $Nd \approx N$, and at the median stream ($N = 352$) it gives $\kappa_s = 4.0$–7.4 against the 1.0 it must clear. We report that rather than assert a wall. What bars $\mathcal{C}_\mathrm{s}$ here is the method:

Proposition 4 (The eviction ceiling)**.** *Let the state maintain uncapped cumulative counters, an episodic log truncated to its last L entries, and any number of further fields none of which is written at index t with a record of $V(x_t)$. On an item whose value sits at index t, $\mathrm{Cap}_B(\mathcal{C}_\mathrm{s}) \leq \Pr[t > N - L] + (1 - \Pr[t > N - L])/K$, while on a state-shaped family whose answer is a counter the ceiling is* 1.

This is a claim about an implementation, not about information, and $L = 20$ is the companion method's own parameter carried over unchanged. Its content is the division of labour that makes the conjunctive family two-channel at all: the state can hold the *locator* (a counter over `open fridge` is never evicted) and cannot hold the *value* at occurrence 158 of 331, the log being 20 deep. Table 1 and Figure 2 evaluate Proposition 1 on the three families of §4, every entry computed before any measurement. The count m presupposes a rule for when two annotated actions are the same event; Appendix B makes it explicit, because our own generator got it wrong once (§6).

3 The Conjunctive Fraction of the Field's Corpora

The instrument classifies a query on the two axes of Definition 1: whether the *value* is a percept at one moment or a reduction, and whether the *locator* is given or must be computed. It touches no frame: a design constraint, and the same one a companion adopts for benchmark validity (Anonymous authors, 2026b), since every corpus here distributes its media under an agreement. It is a rule pass and therefore a *lower* bound: 71% of queries are unclassified and every one counted as non-conjunctive.

Table 1: Where strict super-additivity is available, computed before any measurement. Ceilings from Prop. 4 ($\mathcal{C}_s$), Prop. 3 ($\mathcal{C}_v$ on the percept-valued families) and the horizon wall ($\mathcal{C}_v$ on TRK); the last column evaluates Prop. 1. Chance is 12.5%.

family	n	locator	value	$\mathrm{Cap}(\mathcal{C}_s) \leq$	$\mathrm{Cap}(\mathcal{C}_v) \leq$	$\Sigma - \mathrm{Cap}(\varnothing)$	super-additivity
RET	245	content-findable	percept	18.6%	100.0%	106.1%	○ not certified
TRK	245	—	reduction	100.0%	14.3%	101.8%	○ not certified
CONJ	245	**reduction**	percept	23.6%	31.2%	42.3%	**available**

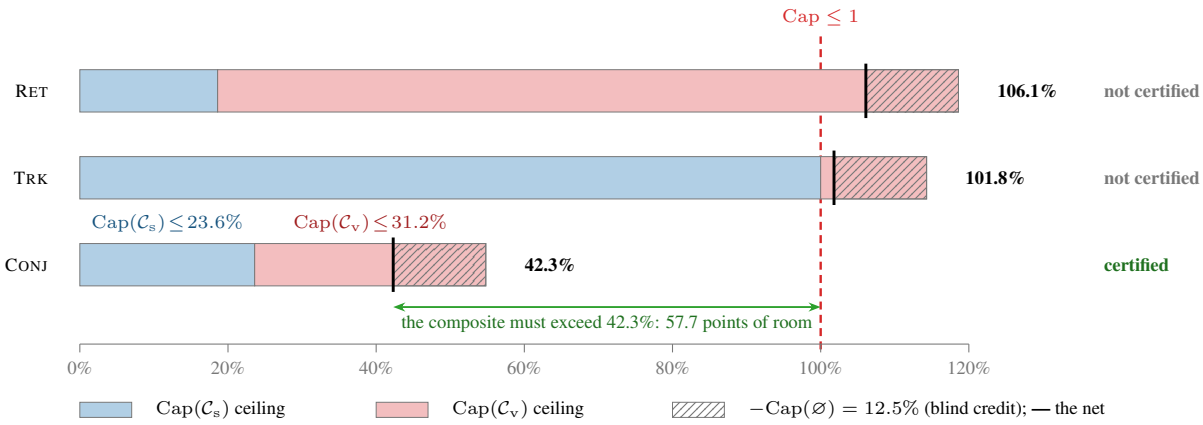


Figure 2: A benchmark can forbid the effect it is being used to look for. Each bar stacks the two single-channel ceilings of Table 1; the hatched tail is the blind credit that Eq. (3) subtracts, so the black tick is the net. Where the net of two *ceilings* stays below 100% the additive inequality is certified available before any run (Prop. 1); where it passes 100% nothing is certified, because a ceiling sum can exceed a measured sum. Only the conjunctive family is certified, and the number it must beat, 42.3%, is derived rather than chosen; the two single-axis families are settled by the measured sums of Table 12. Every value is computed from the corpus and from frozen method parameters before any backbone was run.

Table 2: The conjunctive fraction of six public embodied and video QA corpora, from annotations alone. A rule pass, therefore a *lower* bound: every unclassified query is counted as non-conjunctive.

corpus	questions	conjunctive	p_C
Video-MME	2,700	51	1.89%
PerceptionTest	19,140	127	0.66%
ActivityNet-QA	8,000	29	0.36%
NExT-QA (open)	9,178	23	0.25%
TempCompass	1,580	3	0.19%
VSI-Bench	5,130	0	0.00%
pooled	**45,728**	**233**	$\leq$ **0.51%**
Qasper (text control)	281	0	0.00%

The pooled fraction over 45,728 questions is $p_C \leq 0.51\%$, and the items returned are unmistakably the right shape: *"what colour pen did the author use when he wrote 'guitar' for the second time?"* (Video-MME); *"what object or objects were put in second?"* (PerceptionTest). **VSI-Bench contains none**: its questions are distances, object sizes and egocentric directions, pure spatial state with no percept-valued query at all. It is one of the three natural-video corpora the companion registered and could not obtain; had it obtained them it could not have measured super-additivity on this one, because Proposition 1 forbids it before any licence question arises. **The text control returns zero and is too small to carry weight**: Qasper yields $p_C = 0.00\%$, but at $n = 281$ that zero carries a 95% upper bound of 1.35%. The six corpora were fixed before the instrument was written, by two rules (Appendix C).

3.1 DETECTABILITY: WHY p_C IS A STATEMENT ABOUT INSTRUMENTS

A benchmark's expected composite margin is $\approx p_C\, m_{conj} - (1 - p_C)\, c_{mix}$, where m_{conj} is what a composite can win on a conjunctive item and c_{mix} what it pays on a single-axis one. Bound m_{conj} generously by the whole span from blind to solved, $1 - 1/K = 87.5$ points at $K = 8$, and set $c_{mix} = 0$: both maximally favourable to detection. At the measured fractions, a system that solved the conjunctive subset *perfectly* would move the richest corpus by 1.7 points and the median one by 0.32.

Two noises can hide a margin that small, and they behave differently in n. *Sampling* noise shrinks as $1/\sqrt{n}$, and against it alone these margins are not small ($z \approx 2$–11 on the five corpora with a non-zero fraction). *Run-to-run* noise, redrawing the write phase of the same pipeline on the same corpus under the same prompts and budget, does not shrink with n at all, and is the floor a comparison between two memory systems is actually read against. Writing f for it, detection requires $p_C \geq f/87.5$. Stated parametrically, so nothing turns on one imported constant: a companion measures $f = 9.0$ points of mean $|\Delta|$ this way (Anonymous authors, 2026d), putting break-even at $p_C \geq 10.3\%$ against a measured lower bound of 0.51%, an under-count of 20×; a reader who rejects that number may substitute their own f and read the inequality off Figure 13. So: *at the fractions measured here an access-structure effect is smaller than the run-to-run floor of the pipelines these benchmarks are used to compare, and on VSI-Bench it is identically zero* (95% upper bound 0.07%), and the published nulls for hybrid memory are consistent with the corpora rather than with any claim about the architectures. It is falsifiable twice over (raise p_C above 10.3% on any public corpus, or drive the floor below $p_C \cdot 87.5$ points), and we would welcome either.

4 A CONJUNCTIVE WITNESS ON REAL EGOCENTRIC VIDEO

One keyframe per annotated action segment, at its midpoint, so the stream is exactly the sequence of things that happened. From the 8 longest streams of the EPIC-KITCHENS-100 validation split (291–564 keyframes, median 352, 3,092 keyframes and 735 items) come three families: RET (locator content-findable, value a percept; the index channel's), TRK (a running count; the state's) and CONJ (locator the k-th occurrence, a reduction; value a percept; neither alone). CONJ asks: *"Consider the second time you close the fridge during this session. What did you interact with immediately after that?"*, the value being the noun of the following action, which the model must read off a frame because the annotation is never shown to it.

Selection is deterministic and blind to yield: validation split only, streams filtered by action count, then the longest 8 with ties broken on identifier. Length is the criterion for a stated reason: the horizon ceiling is $w/N + (1 - w/N)/K$, so a short stream makes it vacuous; at these streams it is 13.7–14.9%, where at the first twelve streams by identifier (median $N = 65$) it was 23.3% and the experiment would have measured the corpus. Two leaks are closed by construction and one gate enforced: the occurrence *count* never reaches the prompt ("the second time," never "the second of three"); CONJ distractors are drawn *first* from the values at the other occurrences, making Proposition 3's premise true by construction; and the gold at occurrence k must differ from every other occurrence's value or the item is discarded (129 candidates fail that gate). $K = 8$ options, chance 12.5%, gold is a letter, so there is no judge model, noise or cost; the scorer *counts* every prediction resolving to no option. Appendix P shows the construction and the corpus in full.

5 REGISTERED PREDICTIONS, AND THE GRID

Everything in Table 3 was frozen before a paid call, following the pre-registration methodology of Chen et al. (2026); the record, its two amendments and the criteria that failed are verbatim in Appendix L. Arms are BLIND, STATE-ONLY, RETR-ONLY and ASP (both channels with the access loop), at $B = 4096$, on 3 frozen open backbones spanning 12–31B. The verdict column was computed from the per-question records after the run.

Table 3: Every registered criterion and its verdict. Frozen before any paid call (Appendix L); verdicts written from the per-question records. Chance 12.5%; P3's threshold is Table 1's ceiling sum. *Every prediction about the arms fails*, and §6.1 locates the reason upstream of all of them.

criterion	verdict	what was registered, and what was measured
M1 manipulation	**fails**	BLIND within ± 5 of 12.5% on every family; measured 24.9 (RET) and 2.0 (TRK). Repaired in Amendment A0-A1
M2 instrument	holds	no query-independent arm exceeds the horizon ceiling 13.7–14.9%
M3 Assumption 1	holds	RETR-ONLY on CONJ at or below 31.2%; measured 11.4–15.9
P1 dual ordering	**fails**	holds on 0 of 3 backbones, and *reverses* on all three (§6)
P2 scissors on CONJ	**fails**	≥ 10 points over the better single channel registered; holds on 0 of 3
P3 strict super-additivity	**fails**	$\mathrm{Cap}(\mathrm{ASP}) > 42.3\%$ registered; best ASP on CONJ is 14.7%. F2 does not fire (measured sums 15.9–34.3%), so the family was available and the null is architectural
P4 the contrast	not evaluated	P3 fails on CONJ, so there is no contrast to test; delivered instead by §6.2
P5 scale threshold	**fails**	$\geq 40\%$ registered on the two larger backbones; measured 13.5, 8.6, unscorable
A0-W1 witness validity	**holds**	oracle ledger over wrong address at equal cost: $+14.3$ to $+26.3$ points on all 3 backbones, Holm $p < 10^{-5}$; null on both single-axis families (§6.2)

Table 4: The registered grid. Accuracy at B=4096, chance 12.5%; an unparsed prediction scores 0. A superscript is the number of unparsed predictions of 245; † marks a cell above 5% unparsed, reported and *not* scored.

backbone	B	arm	RET	TRK	CONJ
gemma-3-12b	12	BLIND	24.9	9.0	16.7
gemma-3-12b	12	STATE-ONLY	26.1^{6}	13.5	18.4^{4}
gemma-3-12b	12	RETR-ONLY	16.3	25.3	14.3
gemma-3-12b	12	ASP	20.4	19.2	14.7
qwen3.8-27b	27	BLIND	14.7	2.0	15.1
qwen3.8-27b	27	STATE-ONLY	19.2	8.6	17.6
qwen3.8-27b	27	RETR-ONLY	17.1^{5}	27.8	15.9
qwen3.8-27b	27	ASP	$18.0^{\dagger 19}$	13.1	$16.3^{\dagger 16}$
gemma-4-31b	31	BLIND	$24.5^{\dagger 20}$	$0.4^{\dagger 209}$	$13.1^{\dagger 98}$
gemma-4-31b	31	STATE-ONLY	$22.0^{\dagger 72}$	$5.3^{\dagger 137}$	$12.2^{\dagger 75}$
gemma-4-31b	31	RETR-ONLY	13.5	13.1^{8}	11.4
gemma-4-31b	31	ASP	$18.4^{\dagger 15}$	14.3^{10}	$13.9^{\dagger 23}$

6 RESULTS

The grid ran in full: 36 cells, 735 items each scored by letter. Fifteen cells contain predictions with no option letter, almost all on gemma-4-31b, whose BLIND arm answers a count question with "I am an AI and cannot pick up objects" on 209 of 245 tracking items. An unparsed prediction scores 0, so those cells are floors; every cell above 5% unparsed is marked † and excluded from the verdicts.

Every registered prediction fails, and the failure is informative. The composite is never the best arm on the conjunctive family, the family the corpus was built for and the only one the ceilings certified. CONJ peaks at 18.4% (STATE-ONLY, gemma-3-12b) against the 42.3% threshold, and ASP sits at 13.9–16.3%, within four points of chance. The dual ordering is not merely absent but *reversed* on all three backbones (Fig. 3), and neither reason is about the theory. The state's `action_counts` field increments whenever the single-frame update names an action, and an action spanning several keyframes is named on each (one 354-segment stream reached 216 for one verb–noun pair), so a counter that cannot separate a continuing action from a repeated one answers a count question with a frame count, while retrieval lets the backbone count occurrences in captions it can see. On RET,

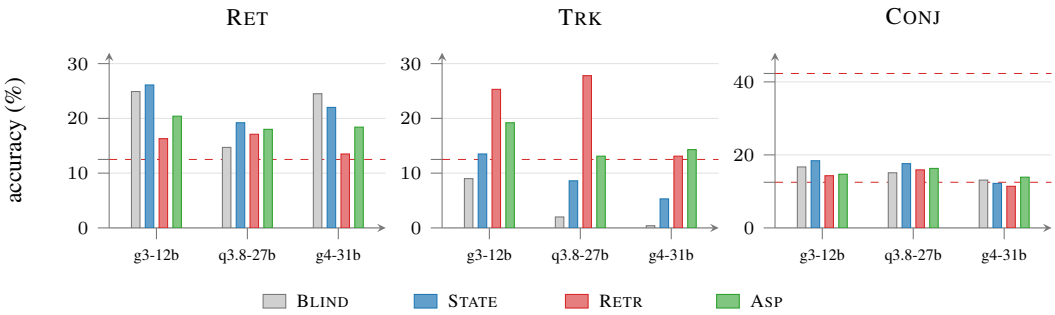


Figure 3: The registered grid, by family. Thirty-six cells at $B = 4096$; dashed lines at chance (12.5%) and, on CONJ, at the availability threshold 42.3%. The composite (ASP) is never the best arm on CONJ, where every arm sits within six points of chance and the threshold is untouched. On TRK the retrieval arm beats the state arm on every backbone, the reverse of the registered dual ordering; on RET the state arm ties BLIND, and BLIND itself is above tolerance on two backbones, the leak §E.2 repairs.

retrieval lands *below* BLIND because the distractor pool leaks to common sense (the M1 check fails there by tolerance, repaired in a registered amendment) and because the retrieved caption names the wrong verb 86–91% of the time (Appendix E).

The measured sums decide what the ceilings could not, and one defect is disclosed. On every family and backbone the measured σ lies between 16 and 35%, nowhere near 1, so the additive inequality was available *everywhere* here and the ceilings' verdict was one-sided exactly as Proposition 1 says: it certified the conjunctive family and was silent, not negative, on the other two. F2 does not fire, so P3 is evaluated, and fails. Separately, the registered generator keyed an occurrence on the raw narration string, where a perceiver can see only the *class*, so a third to a half of the registered items had a false premise. The v3 corpus fixes the definition and is complete at 36 of 36 cells over 308 items; its grid has the registered grid's shape, so the null was not an artefact of the defect (Appendix B).

6.1 UPSTREAM OF EVERY ARM: THE WRITE PATH, MEASURED

The four arms share one write path: a single-frame update naming the action as *verb noun*. Measured on these same cached streams through the corpus's published class tables, that write decides the locator verb correctly 11.0, 15.1 and 23.3% of the time on the three backbones, with errors concentrated on time-reversal pairs: gold *put* written *take* 46% of the time on the most verb-accurate backbone. Read as an index, the event a conjunctive locator names is in the record 12% of the time with its verb and 46% with its noun alone (Figure 14). A retriever cannot address the k-th occurrence of an event the index does not name, and a counter cannot count it, so the null on the conjunctive family is not evidence against composition: it is what the product of an identity factor between 0.11 and 0.23 with anything predicts (Proposition 2(ii), measured).

6.2 WITNESS VALIDITY: THE SAME READ OVER A VALIDATED LEDGER

A null on a corpus built to show an effect admits two readings the grid cannot separate: the corpus permits the effect and the systems could not produce it, or it does not permit it and the ceilings were wrong. A registered amendment fixed the test before its cells were paid for: supply the ledger from the annotations, hold the budgeted read fixed, and ask whether the composed arm clears the threshold. That arm is an address-structured read (Anonymous authors, 2026e): the backbone speaks an address (the k-th occurrence of an event class, $\pm$ an offset), the scaffold resolves it over a ledger of (event, keyframe index) pairs and dereferences exactly one keyframe, and the backbone answers from that frame under the same B=4096 as every other arm. The *oracle* ledger is the corpus's own annotation stream, declared an oracle wherever it appears and never an arm one could deploy; two controls hold it and break the address: *no dereference* picks the frame by similarity, *wrong address* dereferences a different occurrence at the same cost.

That test separates them (Table 16, Appendix P). On the conjunctive family the oracle ledger reaches 31.5% against 13.6% for no-dereference and 17.2% for wrong-address, while on the count family all three sit together at 72–74%; across the three backbones the registered contrast at equal cost is $+14.3$ to $+26.3$ points, Holm $p < 10^{-5}$, 95% episode-bootstrap intervals excluding zero. It is null on both single-axis families and null over the ledger the models themselves wrote (18.8 against 18.5).

Three things follow. The corpus is a valid witness: the availability the ceilings certified is realised, and the dissociation P4 asked for is delivered where the grid could not supply it. The registered grid's loss is located on the write side, by an experiment varying *only* the write. And the wall is not retired (with the ledger made perfect the composite still answers 31.5%, so Proposition 2(iii)'s percept factor is where the remaining loss lives), so we decline to call this strict super-additivity: both single channels score below BLIND here and the additive bound is degenerate.

7 What This Does Not Show

(i) The Shannon wall is vacuous here and we do not get to borrow it ($\kappa_s = 4.0$–7.4 against 1.0), so what bars the compressive channel is an eviction policy (a claim about an implementation, and the one objection a referee is entitled to); the reply is that the round wall, which does the other half of the work, *is* unconditional and Table 1's verdict does not depend on the Shannon wall. **(ii) The conjunctive fraction is a one-sided rule-pass bound over a named population**, 71% unclassified and counted as non-conjunctive; §3.1 survives an under-count of an order of magnitude. **(iii) One corpus, one domain, three backbones, one budget. (iv) The oracle ledger is an oracle**: it prices what a perfect write would buy, and no deployable arm here realises it. **(v) The positive result is a contrast at fixed cost, not a solved family**: 31.5% on CONJ is far from solved, so we have located the wall rather than removed it. Appendix G states all five at length.

8 Related Work

Huh et al. (2024) propose representational convergence; Chen et al. (2026) propose its boundary and name, as its own largest risk, the incidence measurement this paper runs; Anonymous authors (2026d) carries the CCH walls into embodied perception and proves strict super-additivity on a conjunctive witness it then reports it cannot evaluate. Pairing a compressed tier with a retrievable archive is meanwhile the default design for long-horizon agents (Packer et al., 2023; Chhikara et al., 2025), and the same literature reports that the second store often moves little (Wu et al., 2024); distinguishing "hybrid memory does not help" from "the corpus could not have shown it" is what Proposition 1 and §3 are for. The census population is the field's own (Mangalam et al., 2023; Fu et al., 2025; Pătrăucean et al., 2023; Liu et al., 2024; Yang et al., 2025; Xiao et al., 2021; Yu et al., 2019), with OpenEQA (Majumdar et al., 2024) as the embodied comparison; across them the locator is almost always *described* rather than counted. A companion builds the corresponding annotation-only instrument for blind solvability (Anonymous authors, 2026b). Appendix H is the full treatment.

9 Conclusion

CCH names its own largest risk: that the purchase of capability by access structure is never necessary on real work. The answer has two halves pointing opposite ways. The conjunctive load is *rare*: at most 0.51% of 45,728 questions across 6 public corpora, absent entirely from one. It is also *not absent*: it occurs, in the form the theory is about, in corpora nobody in this line of work wrote. Between those facts sits an arithmetic observation that costs nothing: where one channel alone is near ceiling, strict super-additivity is not hard to detect but *unavailable*. Witness-type loads therefore do not have vanishing measure, and have small enough measure that no existing benchmark can see them: a fact about instruments, and a fixable one. On the corpus we build the effect does appear: varying only whether the ledger holds the events the locators name, composition by address pays $+14.3$ to $+26.3$ points and nothing on either single-axis family, but from no deployable arm, because the shared write path does not name the event the locator asks for.

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 L is the registration and its amendments, reproduced verbatim including the criteria that failed. Appendix N proves every proposition stated in the main text. Appendix K gives the prompts verbatim and Appendix I the corpus generator, whose selection is deterministic and blind to yield. The corpus is generated from the public EPIC-KITCHENS-100 validation annotations (Damen et al., 2022) under CC BY-NC 4.0; the generator, the item files, the per-question records for every cell, the analyzers and the annotation-only census instrument are in the supplementary material, and the census reads publicly served question text, gold, category and video identifier only, so a reader can re-run it without holding any corpus's media.

ETHICS STATEMENT

No human subjects were involved and no annotation was collected: the witness corpus is generated mechanically from an existing public dataset's released annotations, used under its licence and for the research purpose it was released for, and the census reads only public question text and gold answers. The backbones are frozen open-weight models called through a commercial router; no model was trained or fine-tuned. The paper's conclusion, that current embodied video benchmarks cannot resolve an access-structure effect, is a limitation statement about evaluation instruments 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 propositions, to formulate the hypotheses, to design the corpus, the census instrument or the experimental grid, to set any pre-registered threshold or falsification criterion, or to interpret the results. Frozen open-weight models are the *subject* of the experiments of §6, not tools used to produce them; every number is computed by an analyzer the authors wrote from the per-question records the runs produced, and the scoring is by option letter with no model judge anywhere in the pipeline. *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 disclosure of the occurrence-semantics defect (§6) and the refusal to call the witness result strict super-additivity (§6.2) 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. The arbitration wall: Reliability-blindness is measurable, and it measures blind. *Manuscript under review*, 2026a. Concurrent anonymous submission; author list withheld under double-blind review. No claim in the present paper depends on it.

Anonymous authors. Two counts over annotations: What a multimodal benchmark can see about its own channels, measured before anything is run, and the refinement that makes the number meaningless. *Manuscript under review*, 2026b. Concurrent anonymous submission; author list withheld under double-blind review. No claim in the present paper depends on it.

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*, 2026c. Concurrent anonymous submission; author list withheld under double-blind review. No claim in the present paper depends on it.

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, 2026d. Concurrent anonymous submission; author list withheld under double-blind review. Cited here for origin and for one empirical constant, the run-to-run floor, which Sec. 3.4 of the present paper treats parametrically so that no conclusion depends on its value.

Anonymous authors. Writing transitions, reading addresses: The write window is the binding parameter of embodied multimodal memory, and an access structure that raises its two walls together. *Manuscript under review*, 2026e. Concurrent anonymous submission; author list withheld under double-blind review. No claim in the present paper depends on it.

Mihir Shriniwas Arya. RECON: Benchmarking agent memory for compositional reasoning over long contexts. *arXiv preprint arXiv:2607.16716*, 2026.

Qirui Chen, Shangzhe Xu, Sifan Yang, and Others. Grounded multi-hop VideoQA in long-form egocentric videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 2159–2167, 2025. arXiv:2408.14469.

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*, 2026. Public preprint, cited in the third person; the framework it states is restated in full wherever this paper relies on it.

Prateek Chhikara, Dev Khant, Saket Aryan, Taranjeet Singh, and Deshraj Yadav. Mem0: Building production-ready AI agents with scalable long-term memory. *arXiv preprint arXiv:2504.19413*, 2025.

Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Evangelos Kazakos, Jian Ma, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision*, 130(1):33–55, 2022. doi: 10.1007/s11263-021-01531-2.

Rosario Forte, Giuseppe Lando, and Antonino Furnari. EGOSTREAM: A diagnostic benchmark for streaming episodic memory in egocentric vision. *arXiv preprint arXiv:2605.31557*, 2026.

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

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

Ningyuan Li, Haiyang Shen, Mugeng Liu, Yudong Han, Zhuofan Shi, and Sixiong Xie. SGR-Bench: Benchmarking search agents on state-gated retrieval. *arXiv preprint arXiv:2605.22219*, 2026.

Yuanxin Liu, Shicheng Li, Yi Liu, et al. TempCompass: Do video LLMs really understand videos? In *Findings of ACL*, 2024.

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

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

Meta AI. S-EMBER: A large-scale benchmark for streaming egocentric memory retrieval. *arXiv preprint arXiv:2607.02689*, 2026.

Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. MemGPT: Towards LLMs as operating systems. *arXiv preprint arXiv:2310.08560*, 2023.

Viorica Pătrăucean, Lucas Smaira, Ankush Gupta, et al. Perception test: A diagnostic benchmark for multimodal video models. In *NeurIPS Datasets and Benchmarks*, 2023.

Toby Perrett, Ahmad Darkhalil, Saptarshi Sinha, Omar Emara, Sam Pollard, Kranti Parida, Kaiting Liu, Prajwal Gatti, Siddhant Bansal, Kevin Flanagan, Jacob Chalk, Zhifan Zhu, Rhodri Guerrier, Fahd Abdelazim, Bin Zhu, Davide Moltisanti, Michael Wray, Hazel Doughty, and Dima Damen. HD-EPIC: A highly-detailed egocentric video dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. arXiv:2502.04144.

Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. LongMemEval: Benchmarking chat assistants on long-term interactive memory. *arXiv preprint arXiv:2410.10813*, 2024.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. NExT-QA: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9777–9786, 2021.

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

Ruoyu Yang and Others. LongVidSearch: An agentic benchmark for multi-hop evidence retrieval planning in long videos. *arXiv preprint arXiv:2603.14468*, 2026.

Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. ActivityNet-QA: A dataset for understanding complex web videos via question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019.

A THE REGISTERED PREDICTIONS, VERBATIM

Table 3 of the main text gives each criterion in one line with its verdict. This is the wording that was frozen, before any paid call, in the registration reproduced in Appendix L. Writing $\sigma(T)$ for the additive quantity of the main text,

$$\sigma(T) \;=\; s + r - b, \qquad s = \mathrm{Cap}_B(\mathcal{C}_\mathrm{s} \,|\, T),\; r = \mathrm{Cap}_B(\mathcal{C}_\mathrm{v} \,|\, T),\; b = \mathrm{Cap}_B(\varnothing \,|\, T), \tag{2}$$

the availability condition of Proposition 1 reads

$$\mathrm{Cap}_B(\mathcal{C}_\mathrm{s} \,|\, T) + \mathrm{Cap}_B(\mathcal{C}_\mathrm{v} \,|\, T) - \mathrm{Cap}_B(\varnothing \,|\, T) \;\geq\; 1 \quad\Longrightarrow\quad \text{no composite is strictly super-additive on } T, \tag{3}$$

and, for a ledger composite with events W (the write resolved the locator) and R (the read returned the gold),

$$\mathrm{Cap}_B(C \,|\, T) \;=\; \Pr[W]\,\Pr[R \,|\, W] \;+\; \Pr[\neg W]\,\Pr[R \,|\, \neg W], \tag{4}$$

which is Proposition 2(i) and the identity every statement below is read against.

Prediction 1 (The dual ordering, on real egocentric video)**.** RETR-ONLY exceeds STATE-ONLY on RET, and STATE-ONLY exceeds RETR-ONLY on TRK, on at least 2 of 3 backbones. This is the companion's central dissociation, which has been shown only on a schematic render.

Prediction 2 (The scissors on the conjunctive family)**.** ASP exceeds the *better* of STATE-ONLY and RETR-ONLY on CONJ by at least 10 points on at least 2 of 3 backbones, paired per question, surviving Holm correction over the tests reported.

Prediction 3 (Strict super-additivity)**.** On CONJ, $\mathrm{Cap}(\textsc{Asp}) - \mathrm{Cap}(\textsc{Retr-only}) > \mathrm{Cap}(\textsc{State-only}) - \mathrm{Cap}(\textsc{Blind})$ on at least 2 of 3 backbones; equivalently, $\mathrm{Cap}(\textsc{Asp})$ exceeds 42.3%, the derived threshold of Table 1. Availability is checked first and reported either way: should the measured single-channel sum reach 1, the family is not conjunctive after all and this prediction is **not evaluated** rather than failed.

Prediction 4 (The contrast that makes the corpus the variable)**.** Prediction 3 holds on CONJ and *fails* on RET and TRK from the same streams, same backbones, same budget. Without this half, Prediction 2 is a fact about a family we chose; with it, the family is the operative variable.

Prediction 5 (A scale threshold, out of corpus)**.** STATE-ONLY on TRK, a count over a $\sim$352-step stream, is at least 40% on the two larger backbones and below 25% on the smallest.

Falsification, as registered. **F1:** STATE-ONLY $\geq$ ASP on CONJ on 2 of 3 backbones: the second channel buys nothing on the family built to require it, and this paper's positive claim fails. **F2:** the measured single-channel sum reaches 1 on CONJ: the generator's gates are wrong. **F3:** BLIND above 25% on any family: corpus leak, run void. **F4:** Prediction 1 fails in both directions: the channel dichotomy does not transfer to real egocentric video.

B Deferred Theory: Measured Bounds, Oracle Witnesses, and Occurrence Semantics

B.1 From ceilings to measurements

The Shannon wall, evaluated. The compressive channel's information bound is $\kappa_s = (M/(Nd) + 1)/\log_2 K$ (Anonymous authors, 2026d), where M is the state's capacity in bits, N the stream length and d the queryable attributes per keyframe. A kitchen annotation carries one queryable attribute per keyframe, so $Nd \approx N$; at the median stream ($N = 352$) and $K = 8$ the wall bites only if $M < 704$ bits, against the 6,144 bits a state of $L_s = 512$ tokens carries. It therefore evaluates to $\kappa_s = 4.0$–7.4 against the 1.0 it must clear, i.e. it is vacuous on this corpus, and §2 reports it as a number rather than asserting a wall. What bars the compressive channel here is the eviction ceiling of Proposition 4, which is a claim about an implementation.

The ceiling sum $\bar{\sigma}(T)$ is what can be computed before a run, and the registered threshold is that number. The quantity Eq. (1) actually compares against is $\sigma(T)$, the measured sum, and the two are related by an identity that fixes which direction the registered number can be wrong in.

Corollary 1 (The measured additive bound)**.** *Fix a family T, a backbone and a budget, and let s, r, b be the measured accuracies of* State-only, Retr-only *and* Blind*, with $\bar{s} \geq s$ and $\bar{r} \geq r$ the ceilings of Table 1. Then:*

(i) a composite C is strictly super-additive on T if and only if $\mathrm{Cap}_B(C \mid T) > \sigma(T) = s + r - b$; this threshold is a function of the three measured single-channel arms and of nothing else;

(ii) T is forbidden in the sense of Proposition 1 if and only if $\sigma(T) \geq 1$, which can be decided only after the three arms have been run;

(iii) the measured threshold and the registered one differ by

$$\bar{\sigma}(T) - \sigma(T) \;=\; (\bar{s} - s) + (\bar{r} - r) + \Big(b - \tfrac{1}{K}\Big), \tag{5}$$

so whenever the blind arm is at or above chance, $\sigma(T) \leq \bar{\sigma}(T)$: clearing the registered threshold implies clearing the measured one, and failing the measured one implies failing the registered one;

(iv) in that case a composite landing in $(\sigma(T), \bar{\sigma}(T)]$ is strictly super-additive while missing the registered number; the registered number is conservative in the one direction that cannot manufacture a positive.

The corollary is bookkeeping, and the point of writing it down is (iii). Prediction 3 registered $\bar{\sigma}$ on the conjunctive family, 42.3%, because it was the only number available before a call; the measured sums of Table 12 are σ, and they are what a composite is actually held to. Each term of Eq. (5) is a slack the ceilings left: the state channel's eviction ceiling assumes the log entry is read perfectly when present; the index channel's ordinal ceiling assumes the retrieval lands some occurrence in the window at all; and the blind arm's nominal $1/K$ is a manipulation check, not a measurement. When all three slacks are non-negative the registered number sits above the true bar, and a composite that misses it may still have been super-additive; a composite that misses the true bar was not, under either number. §F.2 reports both.

B.2 What an oracle witness proves

Proposition 2 names two routes to failure and Eq. (4) shows they enter as a product. A grid of arms cannot separate them, because every arm shares one write map; separating them takes a second ledger over the same streams, written by something whose locator resolution is known. The corpus's own annotations are that something.

Definition 3 (Oracle ledger)**.** Fix an occurrence relation $\sim$ (Definition 4 below) under which the family's items were generated. The *oracle write* ω^* writes, at each keyframe t, the $\sim$-class of the annotated action a_t, and nothing read from the frame. The *oracle composite* is (ω^*, ρ) for the same read loop ρ as the model composite $(\hat{\omega}, \rho)$; the two differ in the ledger and in nothing else. Its write event is W^*, the model composite's is $\hat{W}$.

Assumption 2 (Read equivalence)**.** Conditional on the locator being resolved, the read loop's accuracy does not depend on which write produced the ledger: $\Pr[R \mid \hat{W}] = \Pr[R \mid W^*]$.

The assumption says a resolved ledger is a resolved ledger. It can fail (a model-written ledger that resolves the locator may still surround the right entry with wrong ones that distract the read), and Proposition 5(iii) is stated so that the failure is itself a measurable term rather than a leak.

Proposition 5 (What an oracle witness proves)**.** *On a family T generated under $\sim$, with $A^* = \mathrm{Cap}_B(\omega^*, \rho \,|\, T)$, $\hat{A} = \mathrm{Cap}_B(\hat{\omega}, \rho \,|\, T)$, $\hat{p} = \Pr[\hat{W}]$, $\hat{c}_1 = \Pr[R \mid \hat{W}]$ and $\hat{c}_0 = \Pr[R \mid \neg\hat{W}]$:*

(i) W^* holds on every item, so $A^* = \Pr[R \mid W^*]$ is the read loop's accuracy on resolved ledgers, with no write factor;

(ii) the gap between the two composites decomposes exactly as

$$A^* - \hat{A} \;=\; (1 - \hat{p})\,(A^* - \hat{c}_0) \;+\; \hat{p}\,(A^* - \hat{c}_1), \tag{6}$$

the first term being the mass of unresolved locators weighted by what resolving them is worth to this read loop, the second the read loop's degradation on ledgers that resolve but were written by the model;

(iii) under Assumption 2 the second term vanishes and the write failure rate is identified: $1 - \hat{p} = (A^* - \hat{A})/(A^* - \hat{c}_0)$;

(iv) if $A^* > \sigma(T) \geq \hat{A}$ and $\hat{c}_0 \leq \sigma(T)$, then route (iii) of Proposition 2 is excluded for ρ (the read side clears the measured bound when handed a resolved ledger), and the loss is attributable to the write, with

$$1{-}\hat{p} \;\geq\; \frac{A^* - \sigma(T) + \hat{p}\,(\hat{c}_1 - A^*)}{A^* - \hat{c}_0} \quad \textit{in general, and} \quad 1{-}\hat{p} \;\geq\; \frac{A^* - \sigma(T)}{A^* - \hat{c}_0} \quad \textit{under Assumption 2.}$$

This is what a positive oracle cell buys and what it does not. It does not show that the family is solvable by any frame-reading agent: the oracle reads annotations. It shows that the family is solvable by *this read loop* once the ledger resolves the locator, so that a null from the model composite on the same items with the same read loop is a fact about $\hat{\omega}$, and Eq. (6) says how much of it. The two conditional accuracies in (iv) are not free parameters: $\hat{c}_0$ and $\hat{c}_1$ are scored on the item subsets $\neg\hat{W}$ and $\hat{W}$, which are read off the cached model ledger against the annotations, so every term in Eq. (6) is a measurement. §6.1 reports the write-side quantities the constructive companion measured on these cached streams (Anonymous authors, 2026e); the registered oracle cells (`notes/registered_a0.md`, Amendment A0-A3) are the test of (iv).

The oracle also carries a check the model composite cannot: (i) holds *only if* the oracle writes under the same occurrence relation the items were generated under. An oracle ledger that resolves none of a family's locators is therefore not a broken oracle; it is a proof that the family and the ledger disagree about what an occurrence is. That is how the defect of the next subsection was found.

B.3 What an occurrence is

Every ceiling above, and the locator of every conjunctive item, is built on the phrase "the k-th occurrence of E". The phrase has no meaning until one says when two annotated actions count as the same event, and the corpus gives two candidate answers: the narration strings themselves, and the verb and noun *classes* EPIC-KITCHENS publishes because the strings are not canonical (Damen et al., 2022). The generator's first two versions used the strings. The difference is not a detail of preprocessing; it changes which items exist.

Definition 4 (Occurrence relation, occurrence set, ordinal locator)**.** Let the annotated actions of a stream be $a_{1:N}$, each a string in a set $\mathcal{S}$, and let $\sim$ be an equivalence relation on $\mathcal{S}$ with quotient map $\pi_\sim$. An *event type* under $\sim$ is a class $[E]_\sim$. Its *occurrence set* and *multiplicity* are

$$\mathrm{Occ}_\sim(E) = \{\, t \in [N] : a_t \sim E \,\}, \qquad m_\sim(E) = |\mathrm{Occ}_\sim(E)|,$$

and the *ordinal locator* $L_\sim(E, k)$ is the k-th smallest element of $\mathrm{Occ}_\sim(E)$, defined if and only if $k \leq m_\sim(E)$. The *raw relation* $\sim_{\mathrm{raw}}$ is string equality; the *class relation* $\sim_{\mathrm{cls}}$ holds when two

strings share their verb class and their noun class. Since the class of a string is a function of the string, $\sim_{\mathrm{raw}} \subseteq \sim_{\mathrm{cls}}$: the class relation is the coarser of the two. A conjunctive item under $\sim$ is a triple (E, k, O) with $k \le m_{\sim}(E)$, gold $V(x_{L_{\sim}(E,k)+1})$, and an option set O whose members are pairwise $\sim$-inequivalent; a RET item under $\sim$ requires $m_{\sim}(E) = 1$; a TRK item's gold under $\sim$ is $m_{\sim}(E)$.

Proposition 6 (The locator is defined only relative to the relation)**.** *Let $\sim_1 \subseteq \sim_2$ be two equivalence relations on $\mathcal{S}$, and fix a stream and an event type E. Then:*

(i) $\mathrm{Occ}_{\sim_1}(E) \subseteq \mathrm{Occ}_{\sim_2}(E)$*, hence* $m_{\sim_1}(E) \le m_{\sim_2}(E)$*, and whenever* $L_{\sim_1}(E,k)$ *is defined so is* $L_{\sim_2}(E,k)$*, with* $L_{\sim_2}(E,k) \le L_{\sim_1}(E,k)$*;*

(ii) $L_{\sim_2}(E,k) = L_{\sim_1}(E,k)$ *if and only if no position* $t < L_{\sim_1}(E,k)$ *lies in* $\mathrm{Occ}_{\sim_2}(E) \setminus \mathrm{Occ}_{\sim_1}(E)$*;*

(iii) a RET *item well formed under* $\sim_1$ *is well formed under* $\sim_2$ *if and only if* $m_{\sim_2}(E) = 1$*; a* TRK *item's gold changes if and only if* $\mathrm{Occ}_{\sim_2}(E) \neq \mathrm{Occ}_{\sim_1}(E)$*; a* CONJ *item's locator moves exactly under the condition in (ii), and its option set remains admissible only if no two options are* $\sim_2$*-equivalent;*

(iv) consequently the item set, the gold, the ordinal ceiling $1/m_{\sim}$ *of Proposition 3 and the eviction ceiling of Proposition 4 (through* $t = L_{\sim}(E,k) + 1$*) are all functions of* $\sim$*, and the ceiling sum* $\bar{\sigma}(T)$ *is not defined until* $\sim$ *is.*

Lemma 1 (A class-level perceiver cannot separate class-equivalent options)**.** *Let $\approx$ be the noun-class relation on the value vocabulary $\mathcal{V}$ from which golds and options are drawn: the analogue on values of the relation $\sim$ on action strings, and the relation under which the v3 generator draws its option sets. Call an agent a* class perceiver *if everything it reads from a frame factors through the pair of quotient maps $(\pi_{\sim}, \pi_{\approx})$: it may recover the class of the action shown and the class of the object, and nothing finer. On an item with gold $a \in \mathcal{V}$ and option set O, let $O_a = \{o \in O : o \approx a\}$. Then, even if the perceiver has located the value frame exactly and read its class correctly, its accuracy on the item is at most $1/|O_a|$ under the standing convention; if $|O_a| \ge 2$ the item is unanswerable with certainty by any class perceiver, and the bound is attained by guessing uniformly on O_a. The v1 and v2 option sets were drawn pairwise $\sim_{\mathrm{raw}}$-inequivalent and so admit options with $|O_a| \ge 2$ under $\approx$; the v3 generator draws them pairwise $\approx$-inequivalent, which is the gate this lemma asks for.*

The reading is Amendment A0-A3 of the registered record, and it is the reason there is a corpus v3. Under $\sim_{\mathrm{raw}}$, two narrations of one action are two events: the RET premise "exactly once" is true of a string and false of the event a perceiver sees, the TRK count is a count of spellings, and the CONJ ordinal is an ordinal among spellings. Proposition 6(iii) says exactly which items that breaks and in which direction (coarsening never postpones an occurrence, so a raw-generated "k-th" is at or after the class-level k-th), and Lemma 1 says that an option pool with two class-equivalent entries has put the gold out of reach of any agent whose vocabulary is the class vocabulary, which is what the write path's vocabulary is (§6.1). A model scored against such items fails them for a reason that has nothing to do with access structure, and an oracle ledger written in classes resolves none of them, which is the observation that exposed the defect. The v1 and v2 grids stand as reported; the corrected instrument is generated under $\sim_{\mathrm{cls}}$ with the class-inequivalence gate on options, and every prediction about it was registered before a cell was run.

There is a positive reading as well. Once $\sim$ is fixed, everything in this section is well defined, and the choice is not free: the relation must be one the perceiver can in principle realise, because Lemma 1 makes any finer relation a guarantee of failure. The class relation is the coarsest the corpus publishes and the finest a single-frame write can be asked to name; it is therefore the relation under which a null is a statement about the architecture.

B.4 WHAT THE THEORY PREDICTS AND DOES NOT

The section has one theorem-shaped claim and several bookkeeping results, and it is worth being exact about what follows from them before any cell is read.

It predicts.

1. *Where availability is certified before a run.* Under the stated ceilings, $\bar{\sigma} < 1$ on the conjunctive family and $\bar{\sigma} \geq 1$ on both single-axis families (Table 1); only the first is a certificate, and its threshold is the registered number of Prediction 3.
2. *That the measured threshold is no higher than the registered one* whenever the blind arm is at or above chance (Corollary 1(iii)); and that no family on this corpus is forbidden once the single arms are measured, if their sums are below one, a statement the ceilings could not make.
3. *The ordinal ceiling on* RETR-ONLY. On CONJ, $\mathcal{C}_{\mathrm{v}}$ alone lands at or below $1/m$ averaged over items (Proposition 3); this is M3, and landing far above it refutes Assumption 1, not the proposition.
4. *The eviction ceiling on* STATE-ONLY. On CONJ, $\mathcal{C}_{\mathrm{s}}$ alone lands at or below the value of Proposition 4, which is a function of the log depth and the corpus's locator positions only.
5. *What the oracle settles.* If the oracle composite clears σ and the model composite does not, the read loop is exonerated and the loss is on the write, in the amount Eq. (6) gives (Proposition 5); this is the registered witness prediction of Amendment A0-A3, and it is the one positive result the corrected instrument can deliver independently of any composite's number.
6. *What changing the occurrence relation does.* Regenerating under $\sim_{\mathrm{cls}}$ removes every RET item whose premise was false, changes every TRK gold whose raw and class counts differ, and moves every CONJ locator preceded by a class-only occurrence (Proposition 6); the blind arm on the corrected families should sit at chance because the false-premise items a blind strategy could exploit are gone.

It does not predict.

1. *The value of any composite.* Proposition 2(v) is exactly the statement that $\sigma < 1$ bounds no composite from below. Prediction 3 is a bet about ASP, registered as such, and its failure is a fact about that write map, not a refutation of anything in this section.
2. *Which route binds.* The theory names the write route and the read route and gives each its inequality; which one a given backbone falls through is measured, by Eq. (4) on cached ledgers and by the oracle contrast, and is reported in §6.1 as an imported measurement.
3. *An information-theoretic wall on the state.* The Shannon bound is vacuous here (§7(i)), and Proposition 4 is a statement about a log depth. A state with a different policy has a different ceiling and the availability arithmetic must be redone for it.
4. *Anything about the field's corpora beyond arithmetic.* That single-axis corpora cannot certify access structure before a run is Proposition 1 plus the measured fraction of §3; whether any particular hybrid is good on them is not touched.
5. *Transfer across occurrence relations.* A ceiling computed under one $\sim$ is not a ceiling under another (Proposition 6(iv)); Table 1 is the raw-relation ledger, the class-relation ledger is recomputed rather than inferred (§6), and the two thresholds it yields are 42.3% and 38.1%.

The full proofs are in Appendix N; every one is elementary, and the appendix exists so that each inequality above can be checked against the record without trusting the prose.

C WHAT THE CENSUS COVERS, AND WHAT IT DELIBERATELY DOES NOT

The six corpora were fixed before the instrument was written, by two rules. Their question text, gold, category and video identifier must be served publicly, because an instrument that needs a signed agreement cannot be re-run by a reader (§3). And between them they must cover the four question styles this literature evaluates on: spatial state (VSI-Bench), scripted diagnostic (PerceptionTest, TempCompass), long-form (Video-MME) and web video (NExT-QA, ActivityNet-QA). They include the one natural-video benchmark a companion registered and could not obtain (Anonymous authors, 2026d). That is the population the claim is about, and it is narrower than "video question answering". Three kinds of corpus sit outside it, and each would move p_{C} upward by a knowable mechanism.

Egocentric memory benchmarks built after the registration. A 2025–26 line builds directly for long-horizon egocentric recall: MultiHop-EgoQA pairs multi-hop questions with the several disjoint intervals that answer them (Chen et al., 2025); EGOSTREAM organises streaming episodic-memory

questions along seven cognitive dimensions with an explicit validity window (Forte et al., 2026); S-EMBER poses grounded streaming retrieval over hundreds of hours of wearable video (Meta AI, 2026). These are built for the shape of question our Definition 1 isolates one corner of, and we expect p_{C} on them to exceed anything in Table 2. We have not run the instrument on them, and we do not claim the census speaks for them.

Benchmarks that fix the retrieval interface. LongVidSearch evaluates multi-hop evidence retrieval under a unified tool interface with a metered call budget, and enforces retrieval necessity by construction: a Hop-k question is unsolvable if any one of its k clips is removed (Yang & Others, 2026). That is a sibling of the construction in §4, arrived at independently and for a different reason, and it is the closest existing instrument to the one this paper needed.

Denser annotation on the same domain. HD-EPIC re-annotates kitchen egocentric video far more densely than EPIC-KITCHENS-100 (recipe steps, 3D-grounded object motion, audio), and ships a 26K-question VQA benchmark including multi-hop object-itinerary questions (Perrett et al., 2025). §4 explains why we generate from EPIC-KITCHENS-100 anyway.

The honest statement of the census is therefore conditional and we make it in that form: on the corpora this programme registered and the ones its literature reports on, the conjunctive fraction is at most 0.51%. Whether it is that low on benchmarks written after and against this shape is an open question that the instrument in this section is built to answer, and running it on them is the first item of future work in §7(ii). The availability argument does not depend on the answer: it says a corpus cannot exhibit an effect it does not contain, and a corpus that contains more of it is a better instrument, which is the direction we want the field to move.

D The State Channel Needs the Corpus Ontology

The companion method maintains a typed scene record: cumulative object *sightings*, a door, rooms visited. Run unchanged on EPIC-KITCHENS it produced a state counting sightings (`plate: 29, knife: 39`) and a caption *identical on consecutive frames*: `"dining room, unknown, plate, knife, fork, chair, table, cup, paper"` for five keyframes running. We found this on a two-item smoke run whose only purpose was to read the artefacts, before any grid was paid for.

Both channels die on it. Identical captions give the index nothing to rank on, and the companion measured that captions rather than embeddings carry retrieval on this kind of corpus (recall@8 of 0.936 against 0.005). A sighting counter cannot answer "how many times did you *grab* the salt." State-only would have failed Trk for an implementation reason, which is precisely the mistake the companion's own revision history records and repairs for its synthetic corpus.

So EPIC-KITCHENS gets the same *structure* under its own ontology: one uncapped cumulative reduction over interactions keyed `"<verb> <noun>"`, per-object detail under LRU, a capped episodic log, and a caption that leads with the action. The method's own parameters are unchanged (the log cap stays 20, the LRU width stays 24), and the mapping is read off the corpus's annotation schema, not off any result. After the change, captions on the same episode became `hold cup, cabinet` then `open washing machine, cup, bowl, ...`, and the state carried `"open fridge": 3, "close fridge": 2, "rinse cup": 8`, with zero parse failures over 60 updates.

We report this at length for two reasons. It is the quantity Proposition 4 is about, so a reader cannot check that proposition without it. And it is a transferability result in its own right: an access-structured agent does not port across corpora by changing the data path. Its state schema is a claim about what the world contains, and moving to a world that contains different things requires restating the claim. A paper that transported the method silently would have published an ontology mismatch as a channel result.

E DEFERRED RESULTS

E.1 WHY THE ORDERING REVERSED, READ OFF THE RECORDS

The registered dual ordering was the companion's central dissociation (retrieval wins episodic retrieval, the state wins tracking), and on real egocentric video it reverses on both axes. The records say why, and neither reason is about the theory.

The counter over-counts what persists. The compressive channel's `action_counts` field increments whenever the single-frame update names an action, and an action that spans several keyframes is named on each: one stream of 354 segments reached a count of 216 for a single verb–noun pair. A tracking question asks how many times something happened; a counter that cannot tell a continuing action from a repeated one answers with the number of frames it was visible in. The retrieval arm, by contrast, reads captions that carry the verb–noun pair and lets the backbone count occurrences it can see, which is why it wins the count family on every backbone (25.3, 27.8, 13.1 against 13.5, 8.6 and an unscorable $5.3^{\dagger}$). This is the union-bound decay of the parent system (Anonymous authors, 2026d) in its most literal form, and it is the mechanism the arbitration companion (Anonymous authors, 2026a) measures the cost of.

The retrieval arm cannot see the moment it is asked about. On RET the locator is a description, and the index channel's similarity score should find it. It lands *below* BLIND on two backbones (16.3 against 24.9; 13.5 against 24.5). Two things are happening. The distractor pool leaks to common sense (§E.2), so BLIND is not at chance and the comparison is against an inflated floor. And the retrieved caption names the wrong verb 86–91% of the time (§6.1), so a retriever that finds the right *frame* still hands the backbone a description that contradicts the question. The state arm, holding no queried attribute, answers from the same common sense BLIND does and ties it.

What the reversal does and does not say. It does not say the channel dichotomy is false; it says that on this corpus, under this write path, neither channel holds what its family needs, and what each holds instead (frame counts, wrong verbs) happens to favour the other family. Prediction 1 was a prediction about channels that work; the grid measured channels that share a broken write, and the reversal is a fact about that write.

E.2 THE RET LEAK, ITS CAUSE, AND ITS REPAIR

M1 registered that BLIND lands within ± 5 of chance on every family; it lands at 24.9 and 24.5% on RET for two of three backbones. F3 registered that a BLIND above 25% voids a family; both are below it by a fraction of a point, and we treat the family as leaking rather than hide behind the fraction.

The cause, read off the generator. A RET item asks what the person interacted with immediately after a described moment. Its filler distractors were drawn from the successor nouns of *any* event in the stream, and common sense about the described verb separates them: after "you take the cloth" the successor is not *oven*. The CONJ family never leaked because its sharp distractors are, by construction, the values after the *other occurrences of the same event*, which common sense cannot separate: they are all things one does after opening the cupboard. The leak is therefore not in the corpus's form but in one family's distractor pool.

The repair, registered before its cells. Amendment A0-A1 draws RET's sharp distractors from the successors of the other single-occurrence events with the *same verb* in the same stream, then falls back to the old pool. TRK and CONJ items are byte-identical to v1, their 24 cells carry over, and only the 12 RET cells are re-run into a directory the v1 analysis never reads. The registered prediction: BLIND on RET v2 within ± 5 of 12.5%; if not, the leak is not the distractor pool and the family is reported as leaking with cause unknown. Both versions are reported: v1 as the registered run with M1 failed, v2 as the repair.

What landed, and the registered prediction fails on one backbone and holds on another. All twelve v2 cells are measured (Table 5). BLIND on gemma-3-12b falls from 24.9 to 21.2%: the

Table 5: The RET family before and after the distractor-pool repair (Amendment A0-A1). v1 is the registered run of Table 4; v2 draws the sharp distractors from the successors of the other single-occurrence events with the *same verb* in the same stream. TRK and CONJ items are v1 verbatim and their 24 cells carry over unchanged. Superscripts and † as in Table 4. 12 of 12 v2 cells measured. Registered v2 prediction: BLIND within ±5 of chance: gemma-3-12b FAILS (21.2), qwen3.8-27b HOLDS (16.7), gemma-4-31b unscorable.

backbone	arm	RET v1	RET v2	Δ
gemma-3-12b	BLIND	24.9	21.2	-3.7
gemma-3-12b	STATE-ONLY	26.1[6]	23.7[1]	-2.4
gemma-3-12b	RETR-ONLY	16.3	11.0	-5.3
gemma-3-12b	ASP	20.4	20.4	+0.0
qwen3.8-27b	BLIND	14.7	16.7	+2.0
qwen3.8-27b	STATE-ONLY	19.2	17.1	-2.0
qwen3.8-27b	RETR-ONLY	17.1[5]	14.7[2]	-2.5
qwen3.8-27b	ASP	18.0[†19]	16.7[†20]	-1.2
gemma-4-31b	BLIND	24.5[†20]	19.6[†13]	-4.9
gemma-4-31b	STATE-ONLY	22.0[†72]	20.0[†69]	-2.0
gemma-4-31b	RETR-ONLY	13.5	13.5[1]	+0.0
gemma-4-31b	ASP	18.4[†15]	14.7[†19]	-3.7

direction the cause predicts, below F3's 25% void line, and still 8.7 points above chance against a registered tolerance of 5. By the amendment's own terms the prediction **fails** there: the distractor pool accounts for 3.7 of the 12.4 points above chance and not for the rest. On qwen3.8-27b BLIND moves from 14.7 to 16.7%, *away* from chance, and still inside the ±5 tolerance it was already inside, so the prediction **holds** there without the repair having demonstrated anything; on gemma-4-31b the cell is unscorable by the 5% rule, as its v1 counterpart was. The family is therefore reported as leaking on the smallest backbone with the residual cause open. The candidates the generator leaves are the filler pool that the sharp pool falls back to, and the question stem itself, which names the described moment's verb and noun and thereby a kitchen prior over successors; the v3 corpus (§6) verb-matches the filler pool as well and is the registered next step, not a re-reading of this one. What the twelve cells decide is that the ordering of arms on RET survives an an honest floor: on every scorable backbone the state arm stays within three points of BLIND, the retrieval arm stays below it, and ASP does not move (20.4 → 20.4 on gemma-3-12b). What they cannot decide is any CONJ claim, because CONJ did not change. One cache file was overwritten during the first launch of the repair and re-paid; the deviation is logged with the cells.

E.3 UPSTREAM OF EVERY ARM: THE WRITE PATH, MEASURED

The four arms share one write path: a single-frame update that names the action as *verb noun* and writes it into both the state and the caption index. The constructive companion (Anonymous authors, 2026e) measures that path on these same cached keyframes, and the numbers are what turn the null above into a located result rather than an absence.

The identity factor. At the annotated segment midpoint, the most favourable single frame, the verb is correct 11.0% (gemma-4-31b), 15.1% (gemma-3-12b) and 23.3% (qwen3.8-27b) of the time, against ∼97 verb classes, under a scorer that maps both the written and the gold verb through the corpus's published class tables; the noun overlaps 16–21%. (An earlier version of the companion scored only hyphenated verb heads and reported 6.9, 9.2 and 14.3%; those numbers are retired and appear nowhere here.) The errors are not diffuse. Gold *put* (n=128) is written *take* 46% of the time by the most verb-accurate backbone; gold *close* (n=46) is written *open* 26%; *close* is absent from every backbone's four most frequent writings of itself while *open* is written *open* 31–42%. The confusion is the time reversal, and the companion proves why: a still frame is invariant under reversal of the stream, so no write map over one frame can separate an action from its reverse at any state budget.

What that does to the conjunctive locator. A CONJ item's locator is "the k-th time you *verb noun*." In the single-frame write's own vocabulary that event is findable with its noun alone 46% of the time and with its verb 12–13%. Exact resolution of the k-th occurrence from that vocabulary is

Table 6: The write path every arm shares, measured on this paper's own cached keyframes. At the annotated segment midpoint, the most favourable single frame, *verb* and *noun* are the fraction of segments whose written action matches the gold through EPIC's published class tables, against ∼97 verb classes. The right block reads the written captions as an index and asks what a conjunctive locator could do with it: *findable* is the fraction of CONJ items whose queried event appears at all in the write's own vocabulary, *resolvable* the fraction where it appears at least k times, and *exact* the fraction where the k-th such entry is the annotated locator frame. *pair* is the same verb score when the write sees an ordered pair of frames instead of one. Computed by `a0_stats.py` from `results/axis_write.json` and `results/axis_probe.json`, both released with this paper.

	single-frame write (% of segments)				CONJ locator over that write (%)		
backbone	verb	noun	*put* as *take*	pair verb	findable	resolvable	exact
gemma-3-12b	15.1	16.5	10.9	24.9	19.6	9.8	**1.6**
qwen3.8-27b	23.3	20.8	46.1	31.8	19.2	11.4	**1.2**
gemma-4-31b	11.0	17.5	7.2	22.5	8.2	5.5	**0.5**

0.4–1.6%. A retriever cannot address the k-th occurrence of an event the index does not name, and a counter cannot count it. So the product that a conjunctive query needs (event decided, occurrence resolved, percept read) has a first factor between 0.11 and 0.23 on this write, and every arm inherits it. The CONJ column of Figure 3 is that product, measured from the read side.

What moves when the write is widened, and what does not. The same companion widens the window to an ordered pair of frames under one prompt whose only variation is the attached frames, and verb accuracy at the gold frame rises by 8.6–11.5 points on the three backbones, with the queried event becoming findable 52.7–71.8% of the time. Its registered controls then attribute the rise to the prompt's vocabulary rather than to the window: the same frame attached twice gains as much, the pair in reversed order loses little, and among the *put*/*take* writes the direction is right less often than the majority prior. So the identity factor is not raised by widening the window on these backbones (the ordered pair is presented and not read), and the next factor, per-occurrence detection near 0.2–0.35, binds behind it; the conjunctive family's end-to-end number stays at the floor, and the companion's first end-to-end cells confirm it. That is the companion's result to report. What it settles for this paper is the attribution: the availability condition certified that the effect was permitted on CONJ, the measured sums certified it was permitted everywhere, and the write path is where the permission went unused.

E.4 THE OCCURRENCE SEMANTICS WERE WRONG, AND THE CORRECTED CORPUS

The fourth of our working principles (review the whole definition chain before judging an unexpected number) was triggered not by the grid but by the companion's oracle smoke test (§6.2): a ledger built from the corpus's *own* annotations resolved none of four conjunctive items. An oracle that fails on its own gold is an instrument fault, and the review found it in the first definition of §4.

What the registered generator keyed on. An occurrence was an exact match on the raw narration pair (verb, noun). EPIC-KITCHENS-100 publishes verb and noun *classes* precisely because its narrations are not canonical: "put-down knife" and "put knife" are two events under the raw key and one under the class key; "board:chopping" and "board:cutting" are two options under the raw key and one under the class key. A perceiver, human or model, sees the class. Under class semantics the registered families are wrong by construction on a large fraction of their items: the RET premise "you *verb* the *noun* exactly once" is false on 123 of 245 items because the event recurs under its class; the TRK class count differs from the raw count on 103 of 245; the CONJ ordinal k differs on 80 of 245. Every arm of Table 4, every ceiling of Table 1 and every companion resolution number was scored against these definitions. The registered grid stands as the registered run; what it measured was a family a third of whose locators could not be resolved by anyone.

The ceilings, recomputed under the corrected relation. Because a ceiling is a function of $\sim$ (Proposition 6(iv)), the availability arithmetic of §2 is not inherited: it is recomputed on the v3 items by the same program. The eviction ceiling on the percept-valued families is 22.2%, the ordinal

Table 7: The grid on the class-semantics corpus (v3, Amendment A0-A3). Occurrences keyed on EPIC verb and noun *classes*; 308 items per family; same episodes, keyframes, gates, K=8, seed and budget B=4096 as the registered grid. Superscripts give the unparsed count; † marks a cell above the 5% line, reported and not scored; bold is the best scorable arm in its column and backbone. 36 of 36 cells measured; 11 above the line.

backbone	arm	RET	TRK	CONJ
gemma-3-12b	BLIND	18.5	5.2	**17.2**
gemma-3-12b	STATE-ONLY	**22.4**^{7}	12.0^{2}	16.9^{5}
gemma-3-12b	RETR-ONLY	10.7	**22.4**	12.7
gemma-3-12b	ASP	14.9^{1}	16.6	16.2
qwen3.8-27b	BLIND	17.5	1.6	15.6
qwen3.8-27b	STATE-ONLY	**17.5**	14.3	**18.2**
qwen3.8-27b	RETR-ONLY	15.6^{1}	**28.2**	12.0^{2}
qwen3.8-27b	ASP	18.5^{16†}	11.4	15.6^{26†}
gemma-4-31b	BLIND	16.9^{67†}	0.7^{270†}	9.1^{171†}
gemma-4-31b	STATE-ONLY	23.4^{80†}	4.5^{168†}	11.0^{120†}
gemma-4-31b	RETR-ONLY	**10.4**^{1}	10.7^{16†}	12.3
gemma-4-31b	ASP	15.3^{37†}	15.3^{16†}	**15.6**^{13}

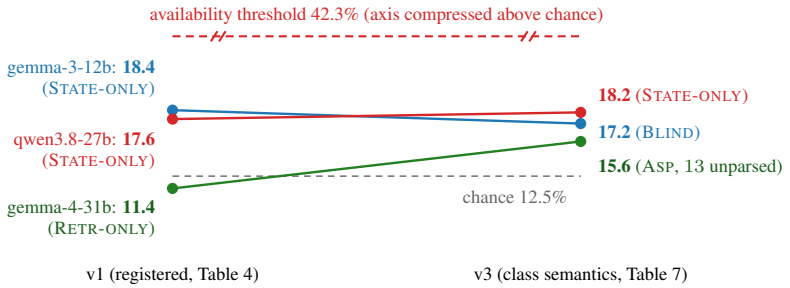


Figure 4: The occurrence-semantics repair did not manufacture the null. Best *scorable* CONJ cell per backbone, registered corpus (v1) against the corrected class-semantics corpus (v3, Amendment A0-A3). On gemma-4-31b every v1 CONJ cell except RETR-ONLY is above the 5% unparsed line, so its v1 point is the one scorable cell; in v3 the completion cap buys a scorable ASP cell. The best scorable arm moves by at most 4.2 points and the threshold is untouched from either side: the dual ordering fails on every backbone in the same direction on both corpora, which makes the reversal a fact about the write path on two corpus versions rather than an artefact of one (§6).

ceiling on CONJ is 28.5%, and the ceiling sum less chance is 109.7% on RET, 101.8% on TRK and 38.1% on CONJ. The verdicts do not change (only the conjunctive family is certified available), but the derived threshold does, from 42.3% under the raw relation to 38.1% under the class relation. Predictions registered against the raw-relation figure are judged against it and against the corrected one, and Table 11 carries both.

The corrected corpus. Amendment A0-A3 (`notes/registered_a0.md`) keys occurrences on (verb class, noun class), writes questions and options in the class vocabulary, draws gold, sharp and filler pools in class keys, verb-matches the RET filler pool as the v2 residual required, and changes nothing else: same eight streams, same keyframes (the frames directory is shared), same gates, same K=8, same seed, same balancing. All three families are regenerated; 308 items per family survive the gates. The v3 predictions were registered before any cell: M1-v3 (BLIND within tolerance on all families, there now being no false-premise items to exploit); P1-v3 and P3-v3 re-tested as registered; and A0-W1-v3, the witness-validity prediction of the next section.

The grid on v3 has the registered grid's shape. Table 7 is the same 36-cell design on the corrected items, complete. The composite is the best scorable arm on the conjunctive family in 1 of the 2 backbones where it is scorable (on gemma-4-31b, where the unparsed rule leaves RETR-ONLY as the only arm to beat and the margin is 3.3 points against a Wilson interval of $[12.0, 20.1]$), so this is not

Table 8: The witness contrast, tested. The registered comparison of Amendment A0-A2 is the composed read over the annotated ledger against the *same* read dereferencing a different occurrence of the same event at the same cost; every row is that pair. *diff* is the paired accuracy difference with a 95% interval from a bootstrap over the eight *episodes*, which prices the possibility that one kitchen carries the effect; n_{10}/n_{01} are the discordant pairs and p is McNemar's exact two-sided test on them, Holm-corrected over the three backbone rows. *sign* counts the episodes in which the oracle leads, with its one-sided sign-test p. The upper block is the conjunctive family on every backbone; the lower block is the same contrast on the two single-axis families, where a witness should show nothing. Unparsed answers score 0 in both arms of a pair, so the marked rows are conservative rather than excluded.

family / backbone	oracle	wrong addr.	diff [95% CI]	n_{10}/n_{01}	p_{Holm}	sign
CONJ, per backbone: the registered contrast						
gemma-3-12b	31.5	17.2	$+14.3$ $[+11.4, +18.7]$	19/63	$<10^{-5}$	8/8, $p = 0.004$
qwen3.8-27b (cap256)†	29.2	12.0	$+17.2$ $[+7.3, +24.9]$	18/71	$<10^{-7}$	7/8, $p = 0.035$
gemma-4-31b (cap256)	37.7	11.4	$+26.3$ $[+22.3, +30.9]$	12/93	$<10^{-15}$	8/8, $p = 0.004$
gemma-3-12b, per family: the dissociation						
RET	20.8	20.1	$+0.6$ $[-5.5, +7.1]$	—	0.88	—
TRK	73.1	74.0	-1.0 $[-2.3, +0.3]$	—	0.38	—
CONJ	31.5	17.2	$+14.3$ $[+11.4, +18.7]$	—	$<10^{-5}$	—
gemma-3-12b, CONJ: the same contrast over the ledger the model wrote						
model ledger	18.8	18.5	$+0.3$ $[-5.3, +7.1]$	—	1.00	—

a result and we do not read it as one. The best scorable CONJ cell on the corpus is 18.2% against the 42.3% threshold; the dual ordering fails on every backbone in the same direction as before, with retrieval winning the count family. The defect therefore did not manufacture the null. The TRK floor is below chance again (5.2 and 1.6% on the scorable backbones), as it already was in the registered grid (9.0 and 2.0%, Table 4); this is not something v3 changed, and the registered results section should have reported it as an M1 failure and did not. A count question offers eight integers and a backbone with no stream picks small ones, so a prior that is systematically wrong scores below a uniform guess. M1 was written for leaks and fails here in the other direction, on both corpus versions; the mechanism is a post-hoc reading and Appendix M gives the answer distribution it rests on. On RET the residual leak persists at 18.5 and 17.5%, five to six points above chance, so verb-matching the filler pool was not the rest of the cause either; the question stem remains the open candidate. gemma-4-31b's BLIND and STATE-ONLY cells are unscorable on every family, as in v1 and v2: the backbone refuses count and first-person state questions rather than picking a letter, and the cells are reported with their counts and not scored.

F THE WITNESS CONTRAST IN FULL, AND THE VERDICT LEDGER

The effect appears, and neither threshold is the right instrument for it. On gemma-3-12b the oracle arm scores 31.5% on CONJ (Table 16), against 17.2% for BLIND and 16.2% for ASP. Against the registered threshold of 42.3%, the raw-relation ceiling sum, prediction A0-W1-v3 **fails**; against the class-relation ceiling sum 38.1% recomputed in §6 it also fails. Both are in Table 11.

It fails on the largest backbone as well. Two further backbones carry a CONJ oracle cell under the completion cap the companion registered for this purpose: gemma-4-31b at 37.7% with 9 of 308 answers unparsed, which is scorable and **fails**; and qwen3.8-27b at 29.2% with 19 unparsed, which is 6.2% and therefore sits above this paper's own 5% line, so it is reported and not scored against the threshold. We state the facts separately because they do different work. The failure against the registered number is a failure on the merits on the two backbones we can score, including the biggest one, rather than a single negative result that a sceptic could attribute to an instrument that could not read its own outputs. What the threshold verdict does *not* settle is the contrast the amendment actually registered beside it, and that contrast is measured on all three backbones; it is the subject of the next two paragraphs.

A reader who has followed Corollary 1 will now reach for the third comparison, the *measured* additive sum $\sigma = \mathrm{Cap}(\mathcal{C}_\mathrm{s}) + \mathrm{Cap}(\mathcal{C}_\mathrm{v}) - \mathrm{Cap}(\varnothing)$, which on these items is 12.3%; the oracle clears it by $+19.2$ points. We report that and decline to call it strict super-additivity, because on this corpus

the criterion is degenerate and the paper's own theory says why. Both single channels here score *below* BLIND (16.9% and 12.7% against 17.2%), so σ falls -4.9 points below the blind floor, and every one of the 10 arms in Table 16 (BLIND included, and both wrong-address controls) clears it. Corollary 1(i) is an *iff* and it is doing its job: it says composition is super-additive exactly when the composite beats what the parts add to, and when the parts subtract, the test carries no information. The standing hypothesis $c_0 \leq \sigma(T)$ that Proposition 2 attaches to its decomposition fails here for the same reason, so Proposition 5(iv) is not available on these cells and we do not invoke it.

What is left when the degenerate comparison is removed is a comparison with a control on the other side, and it is the one the amendment registered: $+14.3$ points over an arm that dereferences a *different occurrence of the same event* at identical cost, and $+14.3$ over blind. That is a statement about the corpus (its conjunctive items are answerable when, and only when, the right occurrence is reached), and it needs no additive arithmetic at all.

The controls say the frame carries the answer, and the contrast is tested on every backbone. The wrong-address control lands at 17.2%, $+14.3$ below the oracle and level with BLIND to a tenth of a point: the same ledger, the same loop, the same cost, and a different occurrence dereferenced, and the whole gain vanishes. The no-dereference control lands at 13.6%, $+17.8$ below: speaking the right address and then choosing the frame by similarity loses more than executing it buys.

Table 8 puts an interval and a test on the comparison, because a $+14.3$-point difference on 308 items that sit in eight kitchens is not a result until both are supplied. The gap is $+14.3$ to $+26.3$ points across the 3 backbones; every backbone's McNemar exact test survives Holm correction over the three at $p < 10^{-5}$; and every backbone's 95% interval from a bootstrap over *episodes*, which prices the possibility that one kitchen carries the whole effect, excludes zero. The distribution-free floor agrees: the oracle leads its control in $8/8$ episodes on gemma-3-12b and $8/8$ on gemma-4-31b, one-sided sign test $p = 0.004$. The two capped rows are marked because their unparsed counts exceed this paper's 5% line, and the mark is conservative rather than exclusionary: an unparsed answer scores 0 in *both* arms of a pair, so it can only shrink the difference.

The dissociation is tested in the same table and is the half that makes it a witness result rather than a difficulty result. On the count family the oracle and its wrong-address control are level (73.0 against 74.0%, -1.0 points, $p = 0.38$; the ledger's count is the answer and the frame is irrelevant), and both are far above every arm of the grid, because a validated event list makes counting a lookup. On RET the oracle sits at 20.8% against 20.1% for the same control ($+0.6$ points, $p = 0.88$) and against a best single-channel arm of 22.4%: an event ledger does not help a description locator, as the companion registered it (its Amendment A5, prediction P-O3). Address execution therefore pays where the locator is a reduction and nowhere else, at $p < 10^{-5}$ on the family that needs it and $p > 0.3$ on both families that do not. That is the pattern a witness should produce, and it is the contrast Prediction 4 asked the registered grid for and could not get.

And the same read over the model's own ledger does not move. The middle block of Table 16 is the identical loop over the ledger the companion's write produced from the same keyframes: 18.8% on CONJ, $+12.7$ below the oracle, within two points of BLIND, and its own wrong-address control at 18.5% beside it, a difference of $+0.3$ points, interval $[-5.3, +7.1]$, $p = 1.00$, which is the null the attribution needs. Breaking the address costs nothing when the ledger does not hold the events, and costs $+14.3$ points when it does. Everything downstream of the ledger is held fixed between the two blocks; the only difference is whether the ledger holds the events the locators name. The registered grid's null is therefore attributed to the write side by an experiment that varies only the write's output, not by the companion's separate measurement of the write's accuracy, though the two agree, and the companion's number says how far the ledger is from the oracle.

Every term of the decomposition, measured (post hoc). This analysis was not registered; it was written after the cells landed and re-reads their traces without adding a cell or an API call, and it is labelled post hoc here and in Table 9. Section B.2 promised that each term of Proposition 5's identity is an observable, and the read traces supply them. Each composed read emits an address; the scaffold logs the keyframe it resolved to; the corpus knows the keyframe the answer sits at. Under the oracle ledger the address reaches that keyframe on 82.8% of conjunctive items, and the read is then correct 32.9% of the time (n=255) against 24.5% when it misses. Over the model-written ledger the same address reaches the value frame on 2.3% of items. So $\Pr[W]$, the write-and-address chain succeeding,

Table 9: Every term of the write/read decomposition, measured. The composed arm emits an address and the scaffold logs which keyframe it resolved to; the corpus knows which keyframe carries the answer. *reaches the value frame* is $\Pr[W]$ of Proposition 2: the write-and-address chain succeeding. The next two columns are c_1 and c_0, the read's accuracy given that it succeeded and given that it did not. The last column is $\Pr[W]c_1 + (1 - \Pr[W])c_0$, which reproduces each cell's measured mean (Table 16) to a tenth of a point. CONJ, gemma-3-12b, v3.

ledger and address	reaches the value frame (%)	c_1	c_0	implied
model-written ledger	2.3	28.6 (7)	18.6	18.8
oracle ledger	82.8	32.9 (255)	24.5	31.5
oracle, wrong address	0.0	—	17.2	17.2
oracle, no dereference	0.6	0.0 (2)	13.7	13.6

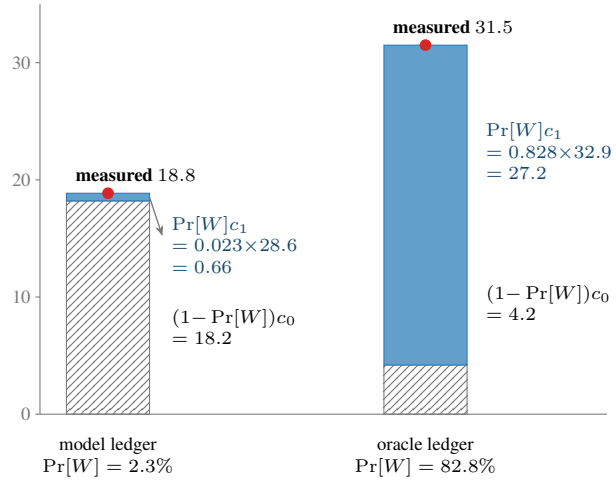


Figure 5: The decomposition is exhaustive, so "the loss is on the write side" is a measured identity. Each bar is the implied decomposition $\Pr[W]c_1 + (1 - \Pr[W])c_0$ (%) from Proposition 2's terms, measured (Table 9); the dot is the arm's measured mean. The write-and-address chain $\Pr[W]$ (the arm's address resolves to the keyframe that carries the answer) moves from 2.3% (model ledger) to 82.8% (oracle), while c_1 and c_0, what the same backbone reads from the right and the wrong frame, stay properties of the read. The stacked sums reproduce each arm's measured mean to a tenth of a point (18.8 and 31.5, dots), which is the check that no unmodelled route carries the result. The controls price c_0 directly: wrong address 17.2, no dereference 13.6.

is 82.8% with the annotations and 2.3% with the model's own record, while c_1 and c_0, everything downstream of the address, are properties of the same backbone reading the same kind of frame. Table 9 gives the terms and shows that $\Pr[W]c_1 + (1 - \Pr[W])c_0$ reproduces every measured mean, which is the check that the decomposition is exhaustive and that no unmodelled route is carrying the result. This is the sense in which the loss is on the write side, stated as a measured identity rather than as an attribution.

Why the count family reaches 73.0% and not 100%. A validated event list makes counting a lookup, and the arm still misses a quarter of the items. The traces say the arm issues no dereference at all on TRK (the address is a count over the ledger, not a frame), so the residual is not addressing but the backbone reading an integer off a correct list and reporting it as one of eight options. An oracle over the queried quantity bounds the read, and the bound is not 1.

Per episode. Table 10 breaks the conjunctive column by stream. The oracle leads its wrong-address control in every stream where the two differ by more than a single item, and the four caption-write arms and the model-ledger arm scatter around chance with no consistent leader. The effect is not carried by one kitchen.

What the oracle does and does not licence. It licences the statement that this corpus is a witness: composition by address pays on its conjunctive family once the ledger is right, and does not pay

Table 10: CONJ accuracy per episode, v3, gemma-3-12b. Integers; n is the episode's CONJ item count. The oracle arm's lead over its wrong-address control is the per-episode form of the gain in Table 16; the four caption-write arms and the model-ledger arm scatter around chance without a consistent leader.

episode	n	BLIND	STATE	RETR	ASP	AXIS	or. wrong	**oracle**
P01_14	52	33	15	17	27	25	25	38
P01_15	18	11	0	6	17	17	6	33
P02_12	22	5	18	5	18	23	27	36
P22_01	31	10	16	16	16	29	10	35
P22_03	24	21	21	12	17	12	17	25
P24_09	65	14	14	12	6	15	11	23
P29_05	38	16	13	13	13	13	18	34
P30_08	58	17	28	12	19	17	21	31

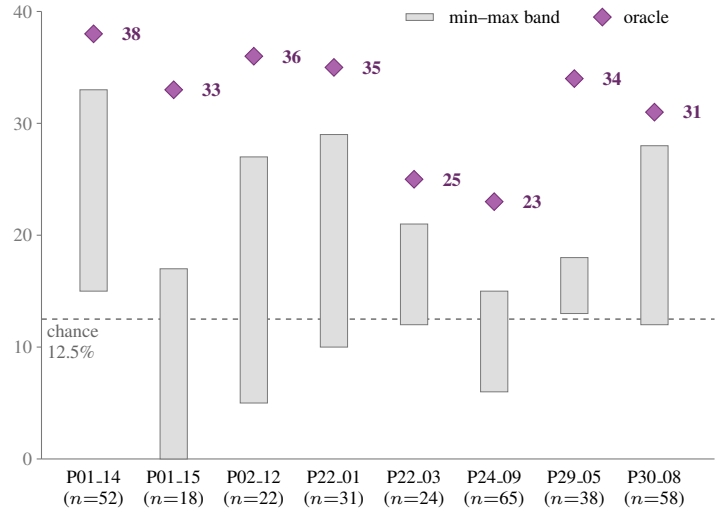


Figure 6: The oracle's lead, per episode (CONJ, v3, gemma-3-12b; post-hoc table 10). The oracle diamond sits above the six non-oracle arms' band on five of eight episodes and inside it on three; unweighted episode means are 31.9 (oracle) against 21.2 (the six arms), and item-weighted the same read is 31.5 against 18.8 (Table 16). The four caption-write arms and the model-ledger arm scatter around chance without a consistent leader, so the gain is carried by the validated ledger and not by the read prompt.

on the two single-axis families, which is the contrast Prediction 4 asked for and the registered grid could not supply. It does not licence any claim about a deployable system, because no system has the oracle ledger; the companion paper owns the question of how close a written ledger can come, and its answer on these backbones is: not close (§6.1). And it prices the percept factor: with the ledger's identity and ordinal content supplied, a single dereferenced frame read under budget answers 31.5% of conjunctive questions. That is not a clean bound on perception alone (the backbone must still name the right event class and ordinal for the scaffold to resolve, so the figure is the product of address formulation with the percept read), but in either reading a perfect write on this corpus would sit far below the ceiling the availability arithmetic promised, on either relation. All three backbones' oracle cells are measured. The registered criterion asked for the *threshold* on two of three, and the threshold fails on every backbone that can be scored against it; the registered *control* contrast succeeds on all three (Table 8). Both verdicts are in Table 11.

F.1 THE AMENDMENTS, JUDGED

Table 11 carries the threshold verdicts and Table 8 the control verdict, and they disagree in a way worth stating plainly. A0-W1 registered two clauses: that the oracle arm clear the ceiling-derived threshold, and that it separate from an equal-cost arm which dereferences the wrong occurrence. The first fails on every backbone the 5% rule lets us score, and we record it as a failure rather than retire the threshold after seeing the number. The second holds on all three, with the magnitude, the

Table 11: Amendments A0-A2 and A0-A3, judged. Criteria verbatim from the registration (Appendix L); values from Tables 7 and 16. The second A0-W1 row is not a registered criterion: it is the same prediction with the ceiling-derived threshold replaced by the measured additive bound of Proposition 1, and is labelled as the post-hoc reading it is.

prediction	backbone	registered criterion	value	verdict
M1-v3	gemma-3-12b	BLIND within ± 5 of chance, all families	18.5 / 5.2 / 17.2	fails, fails, holds
M1-v3	qwen3.8-27b	BLIND within ± 5 of chance, all families	17.5 / 1.6 / 15.6	fails, fails, holds
M1-v3	gemma-4-31b	BLIND within ± 5 of chance, all families	$16.9^{67\dagger}$ / $0.7^{270\dagger}$ / $9.1^{171\dagger}$	unscorable (all)
A0-W1-v3	gemma-3-12b	oracle CONJ > 42.3	31.5	fails
A0-W1-v3	qwen3.8-27b	oracle (cap256) CONJ > 42.3	$29.2^{19\dagger}$	unscorable
A0-W1-v3	gemma-4-31b	oracle CONJ > 42.3	37.7^{8}	fails
A0-W1-v3, measured bound	gemma-3-12b	oracle CONJ $>$ STATE+RETR$-$BLIND	31.5 vs 12.3	holds
P3-v3	gemma-3-12b	ASP CONJ > 42.3	16.2	fails
P3-v3	qwen3.8-27b	ASP CONJ > 42.3	$15.6^{26\dagger}$	unscorable
P3-v3	gemma-4-31b	ASP CONJ > 42.3	15.6^{13}	fails
P1-v3	gemma-3-12b	dual ordering: STATE$>$RETR on TRK, reverse on RET	12.0/22.4; 22.4/10.7	fails
P1-v3	qwen3.8-27b	dual ordering: STATE$>$RETR on TRK, reverse on RET	14.3/28.2; 17.5/15.6	fails
P1-v3	gemma-4-31b	dual ordering: STATE$>$RETR on TRK, reverse on RET	4.5/10.7; 23.4/10.4	fails

correction and the interval in Table 8, and it is the clause that decides whether the corpus is a witness, because it compares two arms that differ only in which frame the address reaches, where the threshold compares an arm against a bound computed under a different occurrence relation from a different corpus version. We report the registered verdict and say which clause we think a reader should weigh.

Table 11 evaluates every prediction registered after the grid. M1-v3 fails on TRK by a floor below chance and on RET by the residual leak, and holds on CONJ on both scorable backbones. A0-W1-v3 fails against the registered raw-relation figure and against the recomputed class-relation one; the measured additive comparison is degenerate on these cells (§6.2) and the table records it as such rather than as a pass. P3-v3 fails where measured, as the write-path result predicted it would. P1-v3 fails in the same direction as P1: the corrected semantics did not restore the dual ordering, which is now a fact about the write path on three corpus versions rather than one.

Three checks from the original registration were never given verdicts in the first version of this paper, and are given them here. **M2** (no query-independent arm above the horizon ceiling, 13.7–14.9%) is **not evaluable**: no uniform-w arm was run, and the two single-channel arms are query-dependent, so the check has no subject. It should not have been registered in that form. **M3** (RETR-ONLY on CONJ at or below Proposition 3's ceiling of 31.2%) **holds** on every backbone and every corpus version, the largest value being 15.9%; Assumption 1 survives. **F1** (STATE-ONLY $\geq$ ASP on CONJ on two of three backbones) **does not fire** on scorable cells: it holds on qwen3.8-27b alone in v1, and an earlier version of this appendix said it fired, which was an error of the same kind the † rule exists to prevent. **F2** (measured single-channel sum reaching 1) and **F3** (BLIND above 25%) **do not fire**; **F4** (Prediction 1 failing in both directions) **fires**, and §E.1 is the reading of it.

F.2 VERDICTS AGAINST THE FROZEN CRITERIA

G WHAT THIS DOES NOT SHOW, IN FULL

(i) The Shannon wall is vacuous here, and we do not get to borrow it. $\kappa_s = 4.0$–7.4 against the 1.0 it must clear. On this corpus the compressive channel is not information-theoretically barred from the conjunctive family; what bars it is an eviction policy (Proposition 4), which is a claim about an implementation. A referee entitled to one objection should make it here. The reply is that the

Table 12: Verdicts against the frozen criteria of §5, computed from the cell records by `analysis/a0_tables.py`. Below: the measured single-channel sum $\mathrm{Cap}(\mathcal{C}_\mathrm{s}) + \mathrm{Cap}(\mathcal{C}_\mathrm{v}) - \mathrm{Cap}(\varnothing)$ per family and backbone: the quantity Proposition 1 actually needs, which the ceilings of Table 1 could only bound from above. Every measured sum is far below 1: no family is arithmetically forbidden on this corpus, and the nulls are architectural.

criterion	verdict	measured
M1 (manipulation check)	FAILS	outside ± 5 of chance: gemma-3-12b/RET 24.9, qwen3.8-27b/TRK 2.0; unscorable cells excluded
P1 (dual ordering)	FAILS	holds on 0 of 3 backbones
P2 (scissors on Conj)	FAILS	holds on 0 of 3
P3 (strict super-additivity)	FAILS	inequality holds on 0 of 3; best ASP on Conj 14.7% against the 42.3% threshold
P4 (the contrast)	NOT EVALUATED	P3 fails on Conj, so there is no contrast to test
P5 (scale threshold)	FAILS	State-only on Trk: gemma-3-12b 13.5, qwen3.8-27b 8.6, gemma-4-31b unscorable
unscorable cells	10 of 36	more than 5% unparsed; gemma-4-31b: ASP CONJ 23, ASP RET 15, BLIND CONJ 98, BLIND RET 20, BLIND TRK 209, STATE-ONLY CONJ 75, STATE-ONLY RET 72, STATE-ONLY TRK 137; qwen3.8-27b: ASP CONJ 16, ASP RET 19

measured sum	gemma-3-12b	qwen3.8-27b	gemma-4-31b
RET	17.6%	21.6%	—
TRK	29.8%	34.3%	—
CONJ	15.9%	18.4%	—

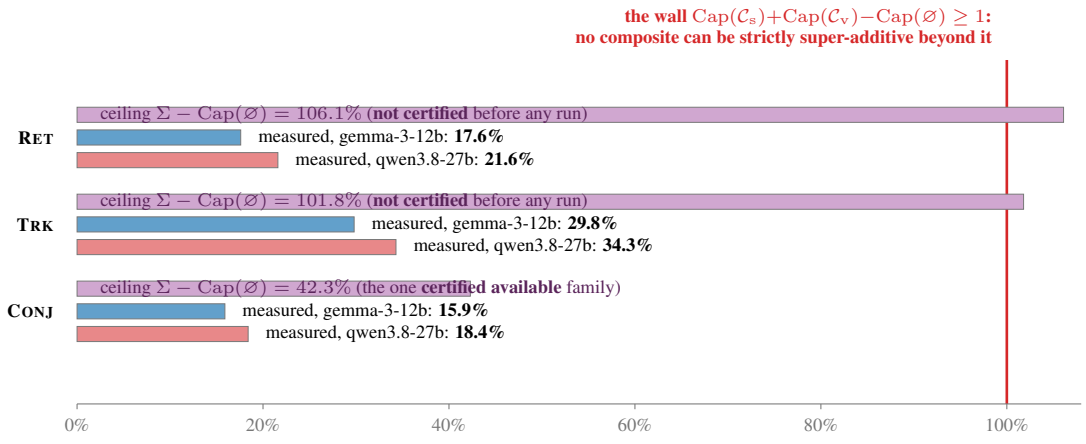


Figure 7: The ceilings certified one family; the measured sums certify all three. The ceiling sums of Table 1 (purple) pass the wall on RET and TRK, so there they certify nothing, and clear it on CONJ, where alone they licence "available". The measured sums of Table 12 (blue/red) lie between 16 and 35% on every family and scorable backbone, nowhere near 1; gemma-4-31b has no measured sum, its cells being unscorable (an em dash in the table). So on this corpus the additive inequality was available everywhere, the ceilings' verdict was one-sided exactly as Proposition 1 says, and the nulls on all three families are architectural, not arithmetical: the binding factor is upstream of the arms (§6.1).

round wall, which does the other half of the work, *is* unconditional, and that Table 1's verdict does not depend on the Shannon wall at all. A corpus whose annotation density made both bounds bite would be a better corpus, and building one is the obvious next step.

(ii) The conjunctive fraction is a lower bound from a rule pass, over a named population. 71% of queries are unclassified and counted as non-conjunctive, so the honest reading of 0.51% is "at most". Two things are outstanding and both are future work rather than gaps in the argument. An adjudicated estimate with an inter-annotator agreement figure would replace the bound with a point estimate; §3.1 is built to survive without it, since the conclusion holds for any p_C below 10.3%, which

is 20× the measured bound. And the census covers the corpora named in §C and not the 2025–26 egocentric-memory benchmarks written for this shape (Chen et al., 2025; Forte et al., 2026; Meta AI, 2026; Yang & Others, 2026) or the denser HD-EPIC question set (Perrett et al., 2025); running the same instrument on those is the measurement we most want next, and we expect it to return a larger number. Nothing in §2 depends on which way it comes out.

(iii) One text corpus is not a comparison, and this one is underpowered. $p_{\mathrm{C}} = 0.00\%$ on Qasper is consistent with CCH's prediction that perception carries more conjunctive load than prose, but at $n = 281$ its 95% upper bound is 1.35%, above four of the six video corpora, so it cannot separate the two regimes at the fractions measured. We report it and draw nothing from it. The comparison worth making (matched domains, matched question-writing protocol, enough text questions to resolve a one-point difference) is future work.

(iv) The witness is generated, and the sense in which that is not circular is specific. We did not sample the conjunctive family from a corpus; we built it. The defence is not that the construction is neutral: it is that §3 independently measures the same form occurring in six corpora nobody involved here wrote, at 0.51%, with examples. Generating more of it is up-sampling a measured tail, not inventing a task. That argument would collapse if p_{C} were zero, which is why the free measurement comes first in the paper and came first in the work.

(v) One activity domain, third-person absent, no action loop. EPIC-KITCHENS is kitchen manipulation from a head camera. There is no navigation, no robot closing its own control loop, and no manipulation of the stream by the agent. This is egocentric perception under a token budget, which is what the theory is about, and it is not embodiment in general.

(vi) Context assembly, not architecture. CCH's classes are architecture families distinguished by state growth and index capacity. Our channels are ways of filling a frozen model's context. CCH licences the move ("an agent's accumulated context is itself a stream, the tool's store an index channel, so the question reappears one level up") and never tests it, so whether the two levels share a capability closure is assumed here and open.

(vii) Nothing here bears on CCH's convergence conjecture. That is CCH's row 9, marked by CCH itself as economics-motivated and consistency-level. Our companion grid gives counter-evidence against it in this regime, with a mechanism. This paper tests the provable core and declines the rest.

(viii) The positive claim of the grid failed, and the located cause is measured on all three backbones. Every registered prediction about the arms failed (Table 12), on three versions of the corpus. The attribution to the write side rests on this paper's own experiment (§6.2): the same read over an oracle ledger separates from its wrong-address control by $+14.3$ to $+26.3$ points on all 3 backbones, Holm-corrected $p < 10^{-5}$ with episode-clustered intervals excluding zero, and over the model's ledger the same contrast is $+0.3$ points ($p = 1.00$). The companion's independent measurement of the write's accuracy agrees. What remains one-backbone-deep is the full per-term decomposition of Table 9, which is post hoc and reported as such.

(ix) One family leaked, and two repairs removed part of the leak. RET v1 fails M1 by tolerance on two backbones; the v2 repair moves the floor 3.7 points toward chance on the smallest backbone, leaving it outside tolerance, and moves the middle backbone 2.0 points the other way from a value that was already inside tolerance; the v3 repair, which also verb-matches the filler pool, leaves 5–6 points above chance on both. The question stem is the remaining candidate and is untested. Every RET comparison in this paper is therefore against a floor a few points too high, and no conclusion drawn from RET depends on the floor being exact. CONJ and TRK are unaffected by construction, and on TRK the v3 floor is below chance, which is a different fact (§6).

(x) Statistical power, and which comparisons it does and does not support. 245 items per family in the registered corpus and 308 in the corrected one, one budget, three backbones. The 95% Wilson half-widths on a cell run 1.1–5.0 points, median 4.0 (Table 13), and the four arms of a CONJ column span at most 6.5 points, so *every* pair of arms on that column has overlapping intervals. The grid

therefore supports one claim (that no arm approaches the threshold) and no ranking among the arms; the paired tests registered in §5 were never reached because no prediction came within range of its threshold. The witness contrast is the one comparison in the paper that is large enough to test, and it is tested rather than asserted (Table 8): paired, Holm-corrected, with intervals from a bootstrap over episodes rather than items.

(xi) The witness is validated with an oracle, and an oracle is not a system. The ledger of §6.2 is the corpus's annotation stream. Its result bounds what composition by address *can* buy on this family and attributes the grid's loss to the write; it says nothing about how a deployable write could close the gap, and the companion's measurement says the gap is wide on these backbones. A reader who wants a gain on a deployable arm will not find one here.

(xii) The registered threshold and the measured bound disagree, and we report both. A0-W1 was registered against the ceilings' sum, computed on the raw-string family before the semantics defect was found. The measured bound is the quantity Proposition 1 is about and is smaller by a factor of three on these items. We did not re-register the threshold after seeing the oracle number, and Table 11 carries the registered verdict (fails) and the post-hoc one (holds) side by side.

(xiii) The occurrence semantics of v1 and v2 were wrong. Section 6 discloses the fraction of registered items with a false premise under class semantics. The grid on the corrected corpus has the same shape, so no conclusion of this paper rested on the defect; the ceilings of Table 1 are recomputed under the class relation in §6, and every threshold in the paper is labelled with the relation it belongs to.

(xiv) There is no query-independent baseline, and M2 has no subject. The registered instrument check M2 asked that no query-independent arm exceed the horizon ceiling, and §F.1 records that it is not evaluable because no uniform-w arm was run: both single-channel arms are query-dependent. A uniform frame-sampling arm at the same budget is the cheapest thing a practitioner would actually run, and its absence means the grid establishes that the composite does not clear the availability threshold without establishing where it sits relative to naive sampling. This does not touch any claim in the paper (the threshold, the oracle contrast and the census are all independent of it), and it is the one additional cell we would add first.

(xv) Three open backbones at 31B and below. The scale range was chosen to straddle a threshold the companion measured, and one of the three (gemma-4-31b) declines to emit an option letter often enough that 11 of its 36 cells are unscorable. Whether the single-frame write path that §6.1 identifies as binding is a property of this scale or of single-frame writing in general is not decided here; the companion's reversal argument says a still frame is invariant under time reversal at *any* budget, which is scale-free, but that is an argument and not this paper's measurement. A frontier-model oracle cell would price it and is future work.

H Related Work in Full

The hypothesis, and the measurement it names as its own risk. Huh et al. (2024) propose that heterogeneous models converge toward a shared representation as they scale; Chen et al. (2026) propose the boundary of that claim: under a fixed per-token budget, representational convergence does not entail capability convergence, and capability converges instead toward architectures holding both a compressive state and a verbatim index. CCH proves three walls and one witness, then names the measurement that would decide whether the witness has measure in natural work and reports that nobody has run it. Our contribution is that measurement, in the perceptual regime CCH does not enter. What CCH takes for granted, and we measure, is that a corpus on which the effect is sought is one on which the effect is arithmetically permitted.

The embodied transport, and the proposition it could not test. Anonymous authors (2026d) carries the CCH walls into embodied multimodal perception as four bounds (a perceptual Shannon wall, a horizon wall, an unconditional round wall and a conditional composition wall), and proves strict super-additivity of its two channels on a conjunctive witness. Its corpus is a synthetic walkthrough

built so that its own bounds are non-vacuous, and it reports in print that the corpus ships no conjunctive subset, so the proposition is not evaluated. It takes for granted that a conjunctive corpus could be built when wanted; §4 is that corpus, and §2 is why its predecessor could not have served.

Composite memory as a system. Pairing a compressed tier with a retrievable archive is now the default design for long-horizon agents: an operating-system analogy pages a summary held in the context window against an archival store reached by function call (Packer et al., 2023); a production memory layer extracts and consolidates facts from conversation and offers a graph-augmented variant beside the vector one (Chhikara et al., 2025). Both are two-channel in CCH's sense, and both are evaluated on conversational corpora whose questions are, by Definition 1, almost entirely single-axis: the locator is supplied and one fetch suffices. They take for granted that a benchmark improvement is the right readout for the second channel; Proposition 1 says the readout is uninformative wherever one channel is near ceiling, and §3 measures how much of the evidence base that describes.

The reported nulls, and what they can mean. The same literature reports that the second store often moves little. The graph-augmented variant of Chhikara et al. (2025) improves on the vector-only one by a few aggregate points; the long-term memory evaluation of Wu et al. (2024) (five abilities over multi-session chat, including temporal reasoning and knowledge updates) finds that value granularity, key expansion and time-aware query rewriting move accuracy, not a second store. Two readings are available: that hybrid memory does not help, or that on a single-axis corpus the additive inequality is unsatisfiable or undetectable and the null says nothing about the architecture. Distinguishing them requires re-running these systems on a conjunctive subset, which our corpus makes possible. The nulls take for granted that their benchmark could have shown the effect; that premise is what we price.

Egocentric long-form video and the reduction-shaped question. EgoSchema draws multi-minute clips from a large egocentric corpus and asks multiple-choice questions generated from dense narrations and filtered by annotators, selecting for a long "certificate": the portion of the clip a human must watch to answer (Mangalam et al., 2023). A long certificate is a reduction (what the wearer was doing overall), which is state-shaped in our vocabulary. Video-MME spans short to multi-hour videos with human-written multiple-choice items across perception, counting, recognition and reasoning (Fu et al., 2025), and is the richest corpus we measure. Both take for granted that "long" is a property of the video; Definition 1 says the property that matters for two channels is whether the locator must be computed, and §3 counts how often it must.

Embodied question answering over scanned interiors. OpenEQA poses open-vocabulary, human-written questions about episodes in scanned and simulated homes, judge-scored, across object recognition, localisation, attributes, spatial understanding, functional reasoning and episodic memory (Majumdar et al., 2024). VSI-Bench assembles spatial questions (counts, distances, object and room sizes, relative directions, route plans and, in one category, the order in which objects appeared) over indoor scans (Yang et al., 2025). Neither, in our reading, contains a question whose locator is a reduction and whose value is a percept: VSI-Bench's appearance-order items ask for the ordering itself, which is a reduction, and its metric items are state. They take for granted that spatial memory is one capability; the availability condition says a spatial-state corpus is one on which an index-shaped channel can never be shown to pay.

Scripted perception and diagnostic control. The Perception Test films crowd-workers acting out scripts designed to exercise memory, abstraction, physics and semantics, with multiple-choice questions beside tracking and grounding tasks (Pătrăucean et al., 2023). Scripting is what makes it diagnostic: because the authors control what happens, they can ask about the second object placed rather than the object placed, which is why it contributes the largest conjunctive count and the second-largest fraction in Table 2. TempCompass isolates temporal aspects (action, speed, direction, order, attribute change) and pairs each clip with a conflicting variant, such as a reversed one, so a static answer is wrong on one of the pair (Liu et al., 2024); its target is the single-frame shortcut that the constructive companion proves is a wall (Anonymous authors, 2026e). Both take for granted that a controlled question set suffices for its target; even a controlled set leaves the two-channel comparison closed if its locators are given rather than computed.

Web video question answering, and the shape of its questions. NExT-QA asks causal, temporal and descriptive questions about everyday web video, in multiple-choice and open-ended form, with temporal items asking what happened before or after a named event (Xiao et al., 2021). ActivityNet-QA asks short open-ended questions in motion, spatial, temporal and free-form templates over untrimmed web video (Yu et al., 2019). Both supply the locator in the question ("before she put it down"), and a content retrieval that finds the named event is one hop from the answer. That is index-shaped, which is why these corpora sit low in Table 2: the rare conjunctive item is one whose anchor is itself an ordinal. They take for granted that a temporal question tests temporal memory; Definition 1 separates a locator that is described from one that must be counted, and only the second obstructs the index.

The corpus we build on, and what its annotation affords. EPIC-KITCHENS-100 is unscripted head-mounted kitchen video with dense narrations, each action segment carrying start and stop timestamps, a verb class and a noun class, released without a signed agreement (Damen et al., 2022). None of its challenges is question answering. What the annotation affords is mechanical gold: the k-th occurrence of a verb–noun pair is a lookup, the value that follows it is a lookup, and neither is shown to the model, which is why the witness can be judge-free. The benchmark takes for granted that its segments are units of recognition; we read them as the sample points of a stream, which only its density permits.

Shortcut solvability and the blind baseline. A literature documents that video and embodied benchmarks are partly solvable without the video, by question-only baselines, single frames and answer priors. The companion instrument computes the blind bound from annotations alone and shows it is one-sided and biased by template granularity (Anonymous authors, 2026b). Ours is the same kind of object on a different axis: blind solvability asks whether a corpus needs the modality; availability asks whether it can distinguish two ways of spending a fixed budget on the modality. The shortcut literature takes for granted that a benchmark requiring the pixels can rank architectures; §2 is the counter-example, computed before any model runs.

Interaction and super-additivity in evaluation methodology. Ablation practice reports the composite against the best single component, and CCH points out that this comparison is neither necessary nor sufficient for a two-channel claim. The additive form of super-additivity is an interaction term, and because accuracy is bounded the term has a ceiling, which Proposition 1 turns into a pre-registration criterion. To our knowledge no memory-systems evaluation reports the single-channel sum before the composite. The methodology takes for granted that any benchmark can host an interaction test; only those whose single-channel ceilings leave headroom can, and Table 1 computes the headroom per family.

Ordinal and occurrence-indexed questions as a probe class. Questions anchored on the k-th occurrence of an event appear across the corpora above as a minority template and are filed under counting or temporal reasoning rather than treated as a class. We treat them as one, because the anchor is not describable to a content retriever. Assumption 1 states this as a measurable claim and §6 measures it. The corpora take for granted that such items are harder instances of counting; the round wall of Anonymous authors (2026d), specialised in Proposition 3, says they are a different kind of question, with an index ceiling that is a number rather than a difficulty.

Witness validity, and the oracle ledger as a positive control. A corpus built to require two channels has to show that it does. The devices here are the standard anti-degeneracy gates of question generation (§4) with a blind arm registered as a manipulation check, and the family on which the check failed by tolerance is reported as failed (§E.2). The analogue in instrument validation is the positive control the companion runs on a corpus with an independently documented defect (Anonymous authors, 2026b). The ledger of §2 plays that role for the witness, saying before any run what each channel can reach if the gates hold. Generated corpora take for granted that their gates suffice; the ledger and the blind arm make sufficiency a measured claim.

Where a field did notice the shape, it built a corpus for it, and that is the scope of our census. The conjunctive shape is not exotic and we do not claim to have named it first. Outside video, two 2026 benchmarks are built for it deliberately. Li et al. (2026) define *state-gated retrieval*: on

specialised data-retrieval sites "answer-bearing evidence becomes accessible only after establishing the correct site-specific retrieval state," so the locator is a state condition that must be computed and the value is then read at it: Definition 1's shape, transported to the web. Their audit of failed trajectories reports that agents usually reach a relevant source and then establish the wrong state, with retrieval-scope drift and criterion mismatch dominating and final answer composition accounting for only 10.3%: a locator failure, not a value failure, which is what Proposition 1 says the conjunctive family should expose. Arya (2026) likewise builds for compositional reasoning over long contexts and reports the strongest non-oracle system at 22.4%.

Inside video, the same is true of a 2025–26 line this paper's registration predates. MultiHop-EgoQA builds egocentric questions that require several disjoint intervals and grounds each answer in them (Chen et al., 2025); EGOSTREAM and S-EMBER build streaming episodic-memory benchmarks with explicit recall horizons (Forte et al., 2026; Meta AI, 2026); LongVidSearch fixes the retrieval backend and meters tool calls so that what is measured is the agent's ability to plan iterative retrieval, and enforces retrieval necessity by removing-any-clip (Yang & Others, 2026). That last construction is the closest published relative of §4's: it reaches, from an agentic-evaluation motive, the same requirement that no single fetch answer the question. We read the convergence as support for Definition 1 being a natural cut rather than a device of ours.

We take all of this as sharpening the census rather than as a counterexample to it, and we state the scope accordingly: the claim of §3 is about the 6 public corpora named in §C, not about benchmarks in general, and that subsection names the corpora it excludes and the direction they would move the number. Read against these, the finding is not that the shape is unknown but that a corpus cannot exhibit an effect it does not contain, whether the omission was deliberate or inherited, and that the benchmarks the composite-memory literature actually reports its nulls on are the ones with the omission. That is the availability argument, and the external evidence that the shape is buildable is the reason the omission is worth reporting: the 0.51% is a property of what these corpora were built to ask, not a limit on what can be asked.

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., 2026) states the hypothesis that capability is bought by access structure under a fixed per-token budget. Five submissions follow, and no two share a population, an instrument and a question:

- a budgeted access-structure grid on a synthetic walkthrough whose outcome is *accuracy*, which proves the conjunctive proposition it then cannot evaluate (Anonymous authors, 2026d);
- an analysis of arbitration between channels whose outcome is *which exit an access loop takes*, computed from round transcripts (Anonymous authors, 2026a);
- a model-free audit of what a benchmark's own annotations can see (an announcement count and a one-sided blind bound), which calls no model at all (Anonymous authors, 2026b);
- a measurement of what it costs to *check* a perceptual claim on a token-metered interface, whose result is that checking costs what looking costs (Anonymous authors, 2026c);
- a constructive paper that makes the write window an architectural parameter, proves that a single-frame write cannot separate a verb from its time reverse, and measures the write path on the cached streams of this paper (Anonymous authors, 2026e).

This paper sits *before* the first: it measures the free parameter (the conjunctive fraction, and the availability it implies) that decides whether the programme's central comparison was ever testable, and supplies the corpus on which it is. It sits *before* the last, which starts from this corpus and this null.

Nothing here rests on an unpublished companion, and here is how to check that. The companions are under review and carry no archival identifier, so a claim of ours that depended on one would be a claim a reader cannot check. We removed those dependencies rather than disclose them. Concretely: Proposition 3 is the round wall of Anonymous authors (2026d) and is *re-proved* here (Appendix N) rather than cited; the method every arm runs is quoted verbatim in Appendix K, includ-

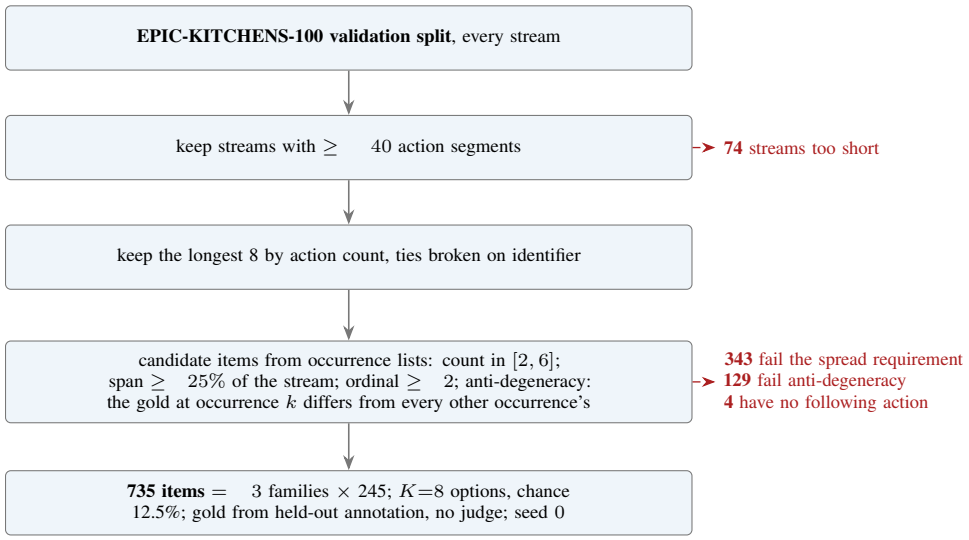


Figure 8: The witness corpus, as generated. One pass, no tuning on yield: selection is deterministic and blind to outcome, and every rejection count is logged. Parameters are quoted verbatim from `analysis/a0_corpus.py` (§I); the gates reject more than they admit, which is what a gate that is doing work looks like, and the anti-degeneracy gate is the companion's blind-baseline lesson moved to the front of the pipeline (§4). Streams and keyframe totals are in Table 15.

ing the address grammar and loop of the composed read introduced by Anonymous authors (2026e); the write-path numbers of §6.1 are computed by a program released with this paper from records released with it (Table 6), not imported; and the one genuinely external constant, the run-to-run floor of Anonymous authors (2026d), enters §3.1 only through an inequality we state parametrically, so the conclusion holds for any floor above a fraction of a point. What the companions are cited for is priority, mechanism and constructive follow-up: Anonymous authors (2026d) for the walls and the dissociation this paper registered a prediction about, Anonymous authors (2026e) for the time-reversal argument and the architecture that answers the null found here, Anonymous authors (2026b) for the annotation-only instrument design whose numbers we share none of, and Anonymous authors (2026a;c) for questions (arbitration and verification price) this paper does not touch.

I THE CORPUS GENERATOR

Frozen parameters, quoted from `analysis/a0_corpus.py`: validation split only; streams with at least 40 action segments; the longest 8 by that count, ties broken on identifier; occurrence counts in $[2, 6]$; occurrences spanning at least 25% of the stream; ordinal at least 2, since "the first time" is the earliest match and a weaker locator; $K = 8$ options; seed 0.

Rejection counts on the validation split: 129 candidates fail anti-degeneracy, 343 fail the spread requirement, 4 have no following action, and 74 streams are too short. The gates reject more than they admit, which is what a gate that is doing work looks like.

J WHY THE AVAILABILITY CONDITION WAS NOT OBVIOUS

Proposition 1 is one line, and the natural question is why a decade of composite memory systems did not stop at it. Three reasons, all visible in the literature.

Papers report *dominance* (composite against the better single channel) rather than the additive inequality, and dominance is available everywhere. CCH's own text warns that better-on-every-task is "trivial and often false," but the weaker comparison is what gets measured, and it does not trigger the arithmetic.

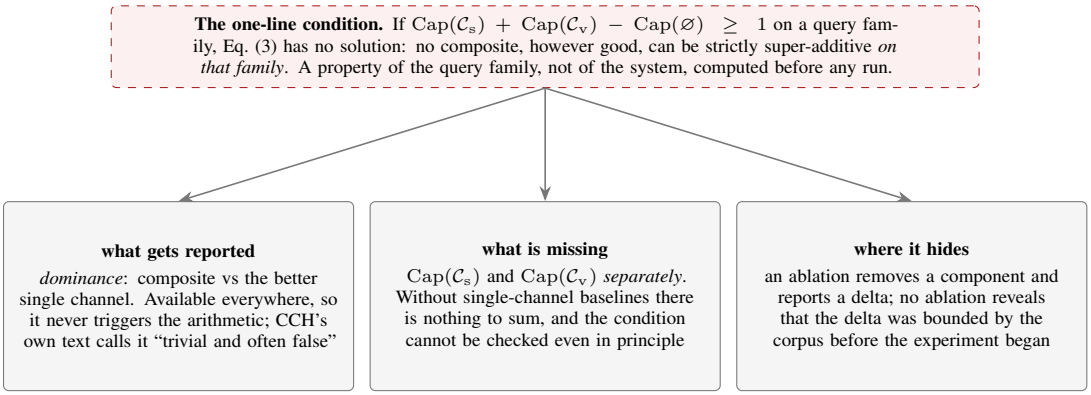


Figure 9: Why the availability condition was not obvious (§J). Three independent reasons, all visible in the literature the condition would have to appear in. The remedy this paper adopts is procedural: the ceilings of Table 1 and the measured sums of Table 12 are computed by the same program, per family, before and after the grid respectively: upper bounds licence "available", only measured sums licence "forbidden".

Single-channel baselines are frequently absent. Without $\mathrm{Cap}(\mathcal{C}_s)$ and $\mathrm{Cap}(\mathcal{C}_v)$ separately there is nothing to sum, so the condition cannot be checked even in principle.

And the condition is a property of the query family, not of the system, so it does not appear in any ablation. An ablation removes a component and reports a delta; no ablation reveals that the delta was bounded above by the corpus before the experiment began.

K PROMPTS, VERBATIM

Every arm of the grid runs the parent system's code path (Anonymous authors, 2026d) with the EPIC ontology bound; the strings below are quoted from `scripts/methods/asp.py` by the generator and are the only instructions any backbone receives. Placeholders in braces are filled by the scaffold.

The write (state update, one call per keyframe). The single-frame update that every arm shares; its `recent_actions` entry is also the caption the retrieval index stores.

```
You maintain a compact JSON interaction state for a person wearing a
head-mounted camera in a kitchen. It is your ONLY memory of the stream: you will
   never see
this frame again, so anything not written down is lost.
Current state (may be truncated): {state}
Given the new keyframe image, output the updated JSON with these fields:
 "action_counts": cumulative count per interaction, keyed "<verb> <noun>" (e.g. "open
   fridge",
   "take plate"). If this frame shows the person interacting with an object,
   increment that
   entry by 1 (create it at 1 if absent). NEVER reset or recount -- these are totals
   over the
   whole stream so far.
 "location": where the person is working now (e.g. "sink", "hob", "fridge",
   "counter").
 "objects": the object classes visible now, each with last_seen={t}.
 "recent_actions": append one entry "t={t}: <verb> <noun>" describing what the person
   does
   with their hands in this frame.
Output ONLY the JSON object.
```

The caption request appended to the update.

```
CAPTION: <what the person is doing with their hands, as '<verb> <noun>', then a
   comma-separated list of the other object class names visible in this frame. Names
   only -- do NOT state their colours.>
```

The router. Used by ASP only; the single-channel arms pay for no classification call.

```
Classify the query for an embodied agent into exactly one class:
RET (asks a visual detail of a specific past moment/object),
TRK (asks a cumulative quantity or final state: counts, open/closed, containment),
CMP (requires chaining multiple dependent spatial/temporal references),
NOW (about the current view only).
Query: {q}
Scene-state summary: {s}
Answer with the class name only.
```

The access loop. One action per round; STATE-ONLY and RETR-ONLY run the same loop with the other channel returning nothing.

```
You are an embodied QA agent with a token budget. You may act in rounds.
Available actions (output exactly one per round):
RETRIEVE: <rewritten visual search query>   -- fetches top matching past keyframes
STATE: <field or 'all'>                     -- reads the compressed scene state
ANSWER: <final answer>
Question: {q}
{context}
Round {r}/{R}. Think briefly, then output one action on the last line.
```

The retrieval read.

```
Question: {q}
Describe ONLY facts relevant to it in these frames.
```

The forced exit. Sent when the budget reserved for the answer is all that remains.

```
Question: {q}
{context}
Give ONLY the final short answer, no explanation.
```

The address grammar and the composed loop (§6.2). Quoted from `scripts/methods/axis.py`. The help text is the only description of the address language any backbone receives; {`vocab`} is the list of event classes the ledger holds, and {`context`} accumulates dereferenced frames.

```
ADDR: nth(<verb> <noun>, k)   -- the k-th time that interaction happened
ADDR: first(<object>=<status>) -- when the object first ENTERED that status,
                                  e.g. first(fridge=open), last(tap=on)
ADDR: first(<verb> <noun>)    -- the earliest time it happened
ADDR: last(<verb> <noun>)     -- the latest time it happened
ADDR: count(<verb> <noun>)    -- how many times it happened
  any of the first three may be followed by +d or -d to step d interactions
  later or earlier, e.g.  ADDR: nth(open cupboard, 5)+1
```

```
You are an embodied QA agent with a token budget. Your memory of the session is an
INTERACTION LEDGER that recorded every interaction and WHEN it happened, plus the
keyframes themselves. You cannot search the keyframes by description; you address them.
Available actions (output exactly one, on the last line):
{addr_help}
STATE: all                     -- read the compressed scene state
ANSWER: <final answer>
Interactions the ledger can address: {vocab}
Question: {q}
{context}
Round {r}/{R}. Think briefly, then output one action on the last line.
```

The *wrong address* control substitutes a different occurrence of the same event, chosen by `crc32` of the address expression so that a resumed run reproduces the same wrong choice; the *no dereference* control resolves the address, reports it as prose, and then selects a frame by similarity as a retrieval arm would. The *oracle ledger* replaces the (event, index) pairs the model wrote with the corpus's annotated action classes at the same indices, and changes nothing else in the loop.

The question and options. Each item is rendered as the question stem followed by eight lettered options in the corpus's vocabulary (class keys in v3), and the scorer (`scripts/epic_score.py`) reads the first option letter in the completion; a completion with no letter is counted unparsed and scored 0. BLIND receives the stem and options and nothing else.

L THE REGISTRATION AND ITS AMENDMENTS

The file `notes/registered_a0.md` is append-only; each block below is its corresponding section in substance, with the verdict the paper reached.

Frozen before any paid call. The corpus generator's parameters (Appendix I); the ceilings and the availability arithmetic of Table 1; four arms at one budget on three backbones; M1–M3 and Predictions 1–5; falsifiers F1–F4; the rule that cells write to a directory no companion analysis reads. *Verdicts:* Table 12 and §F.1: every prediction fails or is not evaluated; M1 fails by tolerance on RET and below tolerance on TRK; M2 is not evaluable and M3 holds; F4 fires and F1–F3 do not.

Amendment A0-A1 (the RET distractor pool). Cause located in the filler pool; sharp distractors redrawn from same-verb single-occurrence successors; twelve RET cells re-run into a new directory; prediction: BLIND within ± 5 of chance. *Verdict:* fails on gemma-3-12b (21.2), holds on qwen3.8-27b (16.7), unscorable on gemma-4-31b (Table 5). A correction is appended to the amendment: an earlier draft's statement that no cell had unparsed predictions was false, and the generator now counts them (Table 14).

Amendment A0-A2 (the witness is validated positively, or it is not a witness). The companion's oracle-ledger arm adopted as this paper's witness-validity test; prediction A0-W1: the composed arm over the annotated ledger clears 42.3% on CONJ on two of three backbones while the single-channel arms stay where they are, and separates from an equal-cost wrong-address control. *Verdict, split by clause and reported both ways.* The threshold clause **fails**: all three backbones are measured and none reaches 42.3% (31.5, 37.7, 29.2), one of them unscorable by the 5% rule. The control clause **holds** on all three, $+14.3$ to $+26.3$ points, Holm-corrected $p < 10^{-5}$, with episode-clustered intervals excluding zero (Table 8). The reading against the measured additive bound (12.3) is labelled post hoc (§6.2, Table 11).

Amendment A0-A3 (occurrence semantics). The raw-string defect and its counts ($123/245$, $103/245$, $80/245$); the class-semantics generator; predictions M1-v3, P1-v3, P3-v3, A0-W1-v3, and a companion prediction that ordinal resolution over the existing pair caches at least doubles under class semantics. *Verdicts:* Table 11; the companion prediction failed (resolution within three frames moved from 2.4 to 1.6% and from 8.2 to 7.5% on the two backbones), which is recorded in the companion's registration as the finding that the model ledger, not the corpus semantics, limited resolution.

What was never registered and is labelled post hoc wherever it appears. The reading of A0-W1 against the measured bound; the below-chance TRK floor as a prior over small counts; the per-episode tables.

M SECONDARY EXPERIMENTS AND SENSITIVITY

THE GRID WITH INTERVALS

UNPARSED PREDICTIONS, EVERY CELL

Table 14 is the count behind every † in the paper. The pattern is one backbone's: gemma-4-31b declines to pick a letter on count questions ("None of the above") and on first-person state questions ("I am an AI and cannot pick up objects") and does so most under BLIND, where nothing in the context tells it the questions are about a video. The 5% rule is applied per cell and never relaxed; the paper draws no conclusion from a marked cell, and the verdicts of Tables 12 and 11 are computed on scorable cells only.

Table 13: The v3 grid with 95% Wilson intervals ($n = 308$ per cell, chance 12.5%). The intervals are the reason §7(x) asks that the CONJ column be read as ordinal and not as a ranking: on every backbone all four arms' intervals overlap, so the grid establishes that nothing clears the threshold and does not establish any ordering among the arms. Marked cells (†) are above the 5% unparsed line and are intervals on a floor.

backbone	arm	RET	TRK	CONJ
gemma-3-12b	BLIND	18.5 [14.6, 23.2]	5.2 [3.2, 8.3]	17.2 [13.4, 21.8]
	STATE-ONLY	22.4 [18.1, 27.4]	12.0 [8.8, 16.1]	16.9 [13.1, 21.5]
	RETR-ONLY	10.7 [7.7, 14.7]	22.4 [18.1, 27.4]	12.7 [9.4, 16.8]
	ASP	14.9 [11.4, 19.3]	16.6 [12.8, 21.1]	16.2 [12.5, 20.8]
qwen3.8-27b	BLIND	17.5 [13.7, 22.2]	1.6 [0.7, 3.7]	15.6 [12.0, 20.1]
	STATE-ONLY	17.5 [13.7, 22.2]	14.3 [10.8, 18.6]	18.2 [14.3, 22.9]
	RETR-ONLY	15.6 [12.0, 20.1]	28.2 [23.5, 33.5]	12.0 [8.8, 16.1]
	ASP	18.5† [14.6, 23.2]	11.4 [8.3, 15.4]	15.6† [12.0, 20.1]
gemma-4-31b	BLIND	16.9† [13.1, 21.5]	0.6† [0.2, 2.3]	9.1† [6.4, 12.8]
	STATE-ONLY	23.4† [19.0, 28.4]	4.5† [2.7, 7.5]	11.0† [8.0, 15.0]
	RETR-ONLY	10.4 [7.5, 14.3]	10.7† [7.7, 14.7]	12.3 [9.1, 16.5]
	ASP	15.3† [11.7, 19.7]	15.3† [11.7, 19.7]	15.6 [12.0, 20.1]

Table 14: Unparsed predictions per cell, all three corpus versions. A prediction is unparsed when no option letter can be read from it; it scores 0. † marks a cell above the 5% line (reporting rule: reported, not scored). v2 regenerated only RET (· elsewhere); an em dash marks a cell not run. Almost every marked cell is gemma-4-31b refusing to answer a count or a state question in the first person. Cells above the line: v1 10, v2 4, v3 11.

		v1			v2			v3		
backbone	arm	R	T	C	R	T	C	R	T	C
gemma-3-12b	BLIND	0	0	0	0	·	·	0	0	0
gemma-3-12b	STATE-ONLY	6	0	4	1	·	·	7	2	5
gemma-3-12b	RETR-ONLY	0	0	0	0	·	·	0	0	0
gemma-3-12b	ASP	0	0	0	0	·	·	1	0	0
qwen3.8-27b	BLIND	0	0	0	0	·	·	0	0	0
qwen3.8-27b	STATE-ONLY	0	0	0	0	·	·	0	0	0
qwen3.8-27b	RETR-ONLY	5	0	0	2	·	·	1	0	2
qwen3.8-27b	ASP	19†	0	16†	20†	·	·	16†	0	26†
gemma-4-31b	BLIND	20†	209†	98†	13†	·	·	67†	270†	171†
gemma-4-31b	STATE-ONLY	72†	137†	75†	69†	·	·	80†	168†	120†
gemma-4-31b	RETR-ONLY	0	8	0	1	·	·	1	16†	0
gemma-4-31b	ASP	15†	10	23†	19†	·	·	37†	16†	13

THE DEFECT'S FOOTPRINT ON THE REGISTERED ITEMS

Under class semantics, 123 of 245 registered RET items assert an event happened exactly once when its class recurs; 103 of 245 TRK golds differ from the class count; 80 of 245 CONJ ordinals differ. The affected CONJ items are the ones where a synonym narration of the queried event precedes the k-th raw occurrence, so the class-level k-th occurrence is earlier than the raw one and the gold successor is a different object. An arm reading the frames could not have resolved those locators as written; the v3 grid's unchanged shape says the other two thirds were not resolved either.

WHAT V3 CHANGED FOR THE CEILINGS

Table 1's ceilings were computed from the raw-string family's occurrence counts, spans and keyframe totals. Under class semantics occurrence counts rise (synonyms merge) and the anti-degeneracy gate admits a different item set: 308 items per family against 245. Recomputing the same ceilings by the same program on the v3 items gives a conjunctive threshold of 38.1% against the registered 42.3%, with the availability verdicts unchanged on all three families. Both thresholds appear in Table 11, each against the corpus version it belongs to. Neither is the quantity Corollary 1 says

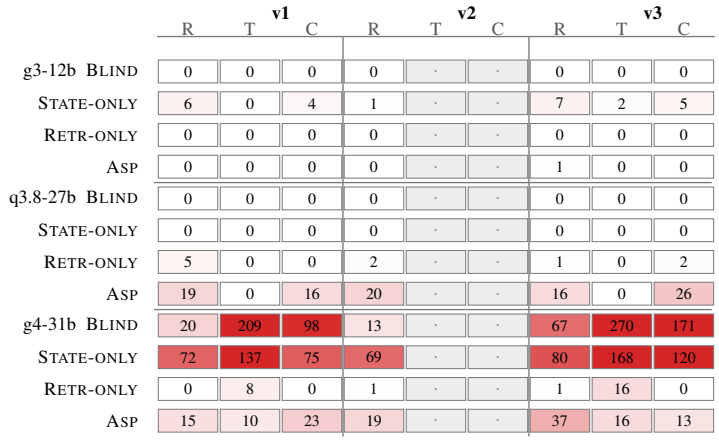


Figure 10: Unparsed predictions per cell (count of 245), all three corpus versions (Table 14; shading scales with the count, · = family not run in v2). Almost every dark cell is one backbone: gemma-4-31b declines to pick a letter on count questions ("None of the above") and on first-person state questions ("I am an AI and cannot pick up objects"), most under BLIND where nothing in the context says the questions are about a video. Cells above the 5% line (13 of 245) are reported and never scored; every verdict in this paper is computed on scorable cells only.

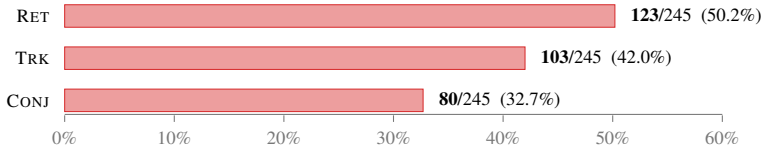


Figure 11: The defect's footprint on the registered items. Bars: the share of the 245 registered items per family whose gold the v1 raw-string semantics made unanswerable by reading: a synonym narration of the queried event precedes the k-th raw occurrence, so the class-level k-th is earlier and the gold successor is a different object (§M). A third to a half of each family's registered golds assert an event happened exactly k times when its *class* recurs at different raw positions; an arm reading the frames could not have resolved those locators as written, which is why the repair was registered before the v3 cells were paid for. v3 re-keys occurrences on EPIC verb–noun classes (Amendment A0-A3) and recomputes the ceilings; the grid's shape is unchanged (Figure 4).

decides super-additivity in a given run; that is the measured additive sum, which on the v3 cells is degenerate (§6.2).

THE RET FLOOR ACROSS THREE CORPUS VERSIONS

BLIND on RET, gemma-3-12b: 24.9 (v1), 21.2 (v2), 18.5 (v3); qwen3.8-27b: 14.7, 16.7, 17.5. Two repairs to the distractor pools removed about half the excess on the backbone that leaks most and none on the one that did not leak. The question stem, which names the described moment's verb and noun, is the remaining candidate, and its test (a stem that withholds the noun where the successor is inferable) is designed and unrun.

THE TRK FLOOR IS BELOW CHANCE, AND WHY THAT IS NOT A LEAK

On v3 the BLIND count arm is at 5.2 and 1.6% against 12.5%. The eight options of a count item are integers, and a backbone with no stream picks small ones; class-level golds are larger than raw ones because synonyms merge. A prior that is systematically wrong scores below a uniform guess. The manipulation check M1-v3 was written for leaks (above tolerance) and fails here in the other direction; the verdict is recorded as written, and the mechanism is the post-hoc reading.

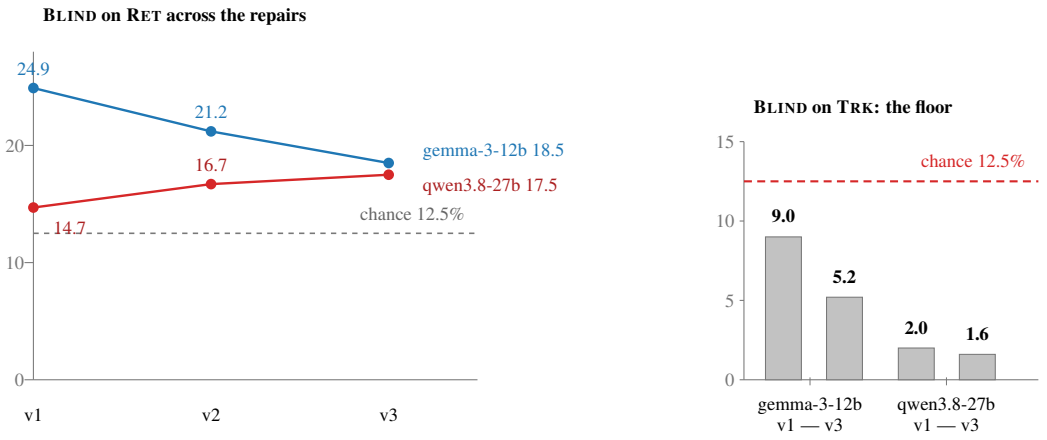


Figure 12: The two floors, and why neither is the result. Left: BLIND on RET (§M): the two distractor-pool repairs (A0-A1, A0-A3) removed about half of gemma-3-12b's excess over chance and none of qwen3.8-27b's; the question stem, which names the described moment's verb and noun, is the remaining candidate, and its test is designed and unrun. **Right:** BLIND on TRK sits *below* chance on both corpus versions: the eight options are integers, a backbone with no stream picks small ones, and class-level golds are larger than raw ones; a prior that is systematically wrong scores below a uniform guess. M1 was written for leaks and fails here in the other direction; the verdict is recorded as written.

N FULL PROOFS

STANDING CONVENTIONS

A query family T is a finite set of items $(x_{1:N}, q, O, a)$ with $|O| = K$ and gold $a \in O$; all probabilities are over the uniform draw of an item from T together with the agent's internal randomness, and $\mathrm{Cap}_B(X \mid T)$ is the probability that arm X returns a. An answer that is not an element of O (an unparsed string, an abstention) is scored as wrong, which is the scoring rule of the grid. On a fixed family b, s, r abbreviate the blind, state-only and retrieval-only accuracies, $\sigma(T) = s + r - b$ is the measured sum of Eq. (2), and $\bar{\sigma}(T) = \bar{s} + \bar{r} - 1/K$ the ceiling sum for any valid upper bounds $\bar{s} \geq s$, $\bar{r} \geq r$. Strict super-additivity of a composite C on T means $\mathrm{Cap}_B(C \mid T) > \sigma(T)$, which is Eq. (1) restricted to T.

Two assumptions are used throughout and were stated in §2; we repeat them so the proofs can point at them. *The gates:* the query text and the option set carry no information about the gold beyond $a \in O$, so an agent whose information about the item, beyond q and O, narrows the gold to a subset $O' \subseteq O$ and nothing finer has success probability at most $1/|O'|$; with $O' = O$ this is the blind rate $1/K$, and it is what M1 tests. Formally: conditional on such an agent's observation, the gold is uniform on O', so any decision rule's success is the expectation of a posterior mass that never exceeds $1/|O'|$. *The percept condition* (Definition 1(ii)): an agent that has read neither the value frame x_t nor any record of it has no information about $V(x_t)$ beyond q and O, and is therefore at $1/K$ on a percept-valued item by the gates.

For a ledger composite $C = (\omega, \rho)$ the events W (locator resolved from the ledger) and R (gold returned) are those of Definition 2; conditional probabilities on an event of probability zero are taken to be zero, which keeps every decomposition below an identity with no division by zero in it.

PROPOSITION 1 (AVAILABILITY)

Proof. Let C be any agent holding both channels. Its accuracy is an item average of indicators, so $0 \leq \mathrm{Cap}_B(C \mid T) \leq 1$. If $\sigma(T) = s + r - b \geq 1$ then $\mathrm{Cap}_B(C \mid T) \leq 1 \leq \sigma(T)$, and Eq. (1) fails for C; since C was arbitrary it fails for every composite. Contrapositively, if some composite satisfies Eq. (1) on T then $\sigma(T) < 1$. The statement is about $\sigma(T)$, which is a property of the three single-channel arms on T and involves no composite, which is why it can be evaluated before one is built, and why, by the same token, it says nothing about any composite once it holds. □

PROPOSITION 2 (AVAILABILITY IS NECESSARY AND NEVER SUFFICIENT)

Proof. Throughout, $p = \Pr[W]$, $c_1 = \Pr[R \mid W]$, $c_0 = \Pr[R \mid \neg W]$, and $0 \le \sigma(T) < 1$.

(i) W and $\neg W$ partition the probability space, so by the law of total probability $\Pr[R] = \Pr[W]\Pr[R \mid W] + \Pr[\neg W]\Pr[R \mid \neg W]$, and $\Pr[R] = \mathrm{Cap}_B(C \mid T)$ by definition of R. This is Eq. (4); it is an identity and assumes nothing about ω or ρ beyond their producing a ledger and an answer.

(ii) Since $c_1 \le 1$, Eq. (4) gives $\mathrm{Cap}_B(C \mid T) \le p + (1-p)\,c_0 = c_0 + p\,(1-c_0)$. The hypothesis $c_0 \le \sigma(T) < 1$ gives $1 - c_0 > 0$, so $c_0 + p(1-c_0) \le \sigma(T)$ is equivalent to $p \le (\sigma(T) - c_0)/(1-c_0)$, which is the hypothesis on p. Hence $\mathrm{Cap}_B(C \mid T) \le \sigma(T)$ and C is not strictly super-additive. The bound used $c_1 \le 1$ and nothing else about the read loop, which is the sense in which the failure is on the write side "whatever the read loop does".

(iii) Eq. (4) writes $\mathrm{Cap}_B(C \mid T)$ as a convex combination of c_1 and c_0 with weights p and $1 - p$, so $\mathrm{Cap}_B(C \mid T) \le \max(c_1, c_0)$. If both $c_1 \le \sigma(T)$ and $c_0 \le \sigma(T)$ then $\mathrm{Cap}_B(C \mid T) \le \sigma(T)$, for every value of p.

(iv) Suppose $c_0 \le \sigma(T)$ and $\mathrm{Cap}_B(C \mid T) > \sigma(T)$. By the convexity bound in (iii), $\max(c_1, c_0) > \sigma(T) \ge c_0$, so $c_1 > \sigma(T)$; in particular $c_1 > c_0$. Substituting into Eq. (4), $\sigma(T) < p\,c_1 + (1-p)\,c_0 = c_0 + p\,(c_1 - c_0)$, and dividing by $c_1 - c_0 > 0$ gives $p > (\sigma(T) - c_0)/(c_1 - c_0)$.

(v) Fix a ledger composite $C = (\omega, \rho)$ and $\alpha \in [0, 1]$. Define the read loop ρ_α that, on each item, draws an independent coin of bias α, runs ρ if the coin lands heads and otherwise returns a fixed string that is not an element of O. Then $C_\alpha = (\omega, \rho_\alpha)$ is a ledger composite with the same write map and the same event W, and, since a non-option is scored wrong and the coin is independent of the item, $\mathrm{Cap}_B(C_\alpha \mid T) = \alpha\,\mathrm{Cap}_B(C \mid T)$. As α ranges over $[0, 1]$ this ranges over $[0, \mathrm{Cap}_B(C \mid T)]$, an interval containing 0. If $\mathrm{Cap}_B(C \mid T) \le \sigma(T)$ take $C' = C$; otherwise take $\alpha = \sigma(T)/\mathrm{Cap}_B(C \mid T) \in [0, 1)$, which is well defined as $\mathrm{Cap}_B(C \mid T) > \sigma(T) \ge 0$, and $C' = C_\alpha$ has $\mathrm{Cap}_B(C' \mid T) = \sigma(T)$. Either way $\mathrm{Cap}_B(C' \mid T) \le \sigma(T)$, so C' is not strictly super-additive, and the set of accuracies attained by ledger composites meets $[0, \sigma(T)]$ (it contains 0 at least). Hence no inequality on $\sigma(T)$ alone can imply that a given composite, or that some composite, is strictly super-additive on T. □

Remark on the hypothesis $c_0 \le \sigma(T)$. It is the one hypothesis in (ii) and (iv) that is not pure bookkeeping, and it is where Definition 1(ii) does its work. On an unresolved item the composite has not identified the value frame; if it also has not read it by some other route, the percept condition puts it at $1/K = b$ on that item, so $c_0 \le b$ up to the composite's occasional luck in retrieving the frame anyway, which is exactly what Proposition 3 bounds. And $b \le \sigma(T)$ whenever $s + r \ge 2b$, that is, whenever each single channel is at least blind. The hypothesis can fail only if the composite answers conjunctive items without resolving their locators, which would mean the items are not conjunctive in its hands; F2 of §5 registers that as a corpus failure, not a finding.

PROPOSITION 3 (AN ORDINAL LOCATOR IS NOT RETRIEVABLE)

Proof. Fix an item whose locator is the j-th of m occurrences $t_1 < \cdots < t_m$ of the event type E, whose gold is $v_j = V(x_{t_j+1})$, and whose option set O contains $v_1, \ldots, v_m$ as distinct entries together with $K - m$ others. Let $J \in [m]$ denote the ordinal named by the query, so the gold is v_J.

What is random, and what is not. J is named by the query, so conditional on q it is *not* random and the bound cannot come from averaging over it. What the gates make exchangeable is the assignment of *values* to occurrences. Write the item as a pair: the observed material $\mathrm{Obs} = (q, O, S)$, and the hidden assignment $\Pi : [m] \to O$ carrying occurrence j to its successor value v_j. The gates require the m values to be distinct members of O and the option set to be drawn so that no value is distinguished by the query; the anti-degeneracy gate removes items whose ordinal is decorative, i.e. exactly those where the assignment is inferable from the query. So the hypothesis we use is exchangeability: *conditional on* (q, O, S)*, the vector* $(v_1, \ldots, v_m)$ *is exchangeable*, so each of the m values is equally likely to be the one sitting at the named occurrence J.

The operative form of Assumption 1. The agent holding $\mathcal{C}_{\mathrm{v}}$ alone observes $\mathrm{Obs} = (q, O, S)$, where S is the retrieved set: at most w frames, each with its content and its stream index, chosen by a rule that ranks frames by a score $s(q, x_t)$ of the query and the frame content. We use the assumption in the form: S carries no information about which occurrence bears which value: formally, the conditional law of S given $(v_1, \ldots, v_m)$ is invariant under permutations of that vector, which is what makes the exchangeability above survive conditioning on S. Assumption 1 as stated is the marginal consequence of this that was measured (under permutation invariance the probability that the true j-th occurrence ranks first among the m is $1/m$ for every j, which is what §4 checks and what `results/a0_assumption1.json` records), and it is the form under which the ordinal is a reduction: the score sees content, the ordinal is not a property of content, and the stream indices returned with the frames tell the agent *where* a retrieved frame is but not *how many* occurrences precede it, which is a count over frames it did not retrieve. The only way the indices would break the invariance is if all m occurrences were returned and the agent could certify that no other occurrence exists, which requires $w \geq m$ together with knowledge of m, itself a reduction the agent does not hold.

The bound. Let δ be any decision rule, a possibly randomised map from Obs to O; note δ may use J, since $J \in q \subseteq \mathrm{Obs}$. Fix Obs and let $A = \delta(\mathrm{Obs})$ be the option it returns, which is now a constant. By exchangeability, conditional on Obs the value at occurrence J is uniform over the m distinct values $\{v_1, \ldots, v_m\}$ in a sense that does not depend on which index J is, so

$$\Pr[\, A = v_J \mid \mathrm{Obs}\,] \;=\; \frac{|\{j : v_j = A\}|}{m} \;\leq\; \frac{1}{m},$$

the inequality because the v_j are distinct, so at most one of them equals A. Averaging over Obs gives $\Pr[\delta(\mathrm{Obs}) = v_J] \leq 1/m$ for every δ, and $\mathrm{Cap}_B(\mathcal{C}_{\mathrm{v}}) \leq 1/m$ on the item. Averaging over the family gives the item-averaged ceiling reported in Table 1. The $K - m$ options outside $\{v_1, \ldots, v_m\}$ are never the gold, so any mass δ places on them is lost; this is the sense in which "options beyond the m are strictly worse". □

Remark. When $m = 1$ the bound is 1, so the proposition is not a bound on retrieval; it is a bound on retrieval *addressed by content* on a locator that content does not address. The dissociation in the text (the same proposition bars $\mathcal{C}_{\mathrm{v}}$ on CONJ and permits it on RET) is the statement that m is a property of the item and not of the channel. Proposition 6(iv) adds that m is a property of the item *under a fixed occurrence relation*, so the ceiling moves when the relation does.

PROPOSITION 4 (THE EVICTION CEILING)

Proof. The state at query time consists of two parts: a set of cumulative counters, each a function of the multiset of event types written so far, and an episodic log holding the entries written at stream indices $N - L + 1, \ldots, N$ and nothing written earlier, because the log is truncated to its last L entries after every update and entries are written once per keyframe without lookahead. Let A be the event $\{t > N - L\}$ that the value frame's entry survives in the log.

On A the state may hold a record of x_t and we bound the agent's success by 1. On $\neg A$ the entry for x_t has been evicted; the state's remaining content is the counters, which are functions of event-type counts and carry no record of the percept at index t, together with log entries for indices other than t, none of which is a record of x_t. The agent holding $\mathcal{C}_{\mathrm{s}}$ alone re-reads no frame. It has therefore read neither x_t nor any record of it, and by the percept condition and the gates its success on $\neg A$ is at most $1/K$. By total probability,

$$\mathrm{Cap}_B(\mathcal{C}_{\mathrm{s}}) \;\leq\; \Pr[A] \cdot 1 + \Pr[\neg A] \cdot \frac{1}{K} \;=\; \Pr[\, t > N - L\,] + \big(1 - \Pr[\, t > N - L\,]\big)\frac{1}{K},$$

where the probability is over the family's distribution of value positions t, which is fixed by the corpus and the generator and so is computable before any run.

For a state-shaped family whose answer is a counter value, the counters are uncapped and are updated at every keyframe, so the counter for the queried event type equals the true count at query time; an agent that reads the counter off is correct on every item for which the counter is right, and 1 is an upper bound on any accuracy, so the ceiling is 1. It is a ceiling and not a prediction: §E.1 measures the same counter over-counting a persisting action by a factor of several, so nothing here says the ceiling is attained, and the grid shows it is not. □

Remark. The bound is exact for the policy and is a statement about L, not about M: doubling the state's token budget while keeping the truncation rule leaves the ceiling where it is, and this is what "the Shannon wall biting in practice on a corpus where it is formally vacuous" means. A policy that indexed the log by event type rather than by recency would have a different ceiling, and Proposition 4 says nothing about it; §7(i) is the admission.

COROLLARY 1 (THE MEASURED ADDITIVE BOUND)

Proof. (i) Eq. (1) restricted to T reads $\mathrm{Cap}_B(C \mid T) > \mathrm{Cap}_B(\mathcal{C}_\mathrm{s} \mid T) + \mathrm{Cap}_B(\mathcal{C}_\mathrm{v} \mid T) - \mathrm{Cap}_B(\varnothing \mid T)$, and the right-hand side is $s + r - b = \sigma(T)$ by definition, so strict super-additivity is $\mathrm{Cap}_B(C \mid T) > \sigma(T)$. The right-hand side involves the three single-channel arms on T and no other quantity.

(ii) This is Proposition 1 restated with the measured $\sigma(T)$ in place of its left-hand side. Since s, r, b are accuracies of arms that must be run to be known, $\sigma(T)$ is not available before the run; a ceiling sum $\bar{\sigma}(T) \geq \sigma(T)$ below 1 certifies $\sigma(T) < 1$, but $\bar{\sigma}(T) \geq 1$ is consistent with either $\sigma(T) \geq 1$ or $\sigma(T) < 1$, which is the one-sidedness stated in §2.

(iii) Subtracting the definitions, $\bar{\sigma}(T) - \sigma(T) = (\bar{s} + \bar{r} - 1/K) - (s + r - b) = (\bar{s} - s) + (\bar{r} - r) + (b - 1/K)$, which is Eq. (5). The first two terms are non-negative because $\bar{s}$, $\bar{r}$ are upper bounds; the third is non-negative if and only if $b \geq 1/K$. Under that condition $\sigma(T) \leq \bar{\sigma}(T)$, so $\mathrm{Cap}_B(C \mid T) > \bar{\sigma}(T)$ implies $\mathrm{Cap}_B(C \mid T) > \sigma(T)$, and $\mathrm{Cap}_B(C \mid T) \leq \sigma(T)$ implies $\mathrm{Cap}_B(C \mid T) \leq \bar{\sigma}(T)$.

(iv) If $\sigma(T) < \mathrm{Cap}_B(C \mid T) \leq \bar{\sigma}(T)$ then C is strictly super-additive by (i) and does not exceed the registered number. Such a C is misreported as a failure by the registered threshold; a C with $\mathrm{Cap}_B(C \mid T) \leq \sigma(T)$ is correctly reported as a failure by both. The registered number can therefore withhold a positive but cannot grant one that the measured bound denies, which is the stated direction of conservatism. □

Remark. If the blind arm falls below $1/K$ the third term of Eq. (5) is negative and the ordering in (iii) can reverse when it outweighs the first two. The manipulation check M1 is two-sided for this reason: a blind arm far below chance is as much a defect in the instrument as one far above it, and §F.2 reports the measured sums rather than inferring them.

PROPOSITION 5 (WHAT AN ORACLE WITNESS PROVES)

Proof. (i) The family was generated under $\sim$, so each item's locator is $L_\sim(E, k)$, the k-th smallest index t with $a_t \sim E$ (Definition 4). The oracle write records $\pi_\sim(a_t)$ at every keyframe t, so the oracle ledger's entries equivalent to E sit at exactly the indices of $\mathrm{Occ}_\sim(E)$, in stream order; the resolver that returns the k-th such entry returns $L_\sim(E, k)$. Hence W^* holds on every item and $\Pr[W^*] = 1$. Applying Eq. (4) to the oracle composite, $A^* = 1 \cdot \Pr[R \mid W^*] + 0 = \Pr[R \mid W^*]$.

(ii) Apply Eq. (4) to the model composite: $\hat{A} = \hat{p}\,\hat{c}_1 + (1 - \hat{p})\,\hat{c}_0$. Then

$$A^* - \hat{A} \;=\; A^*\big(\hat{p} + (1 - \hat{p})\big) - \hat{p}\,\hat{c}_1 - (1 - \hat{p})\,\hat{c}_0 \;=\; (1 - \hat{p})\,(A^* - \hat{c}_0) + \hat{p}\,(A^* - \hat{c}_1),$$

which is Eq. (6). It is an identity.

(iii) Assumption 2 says $\hat{c}_1 = \Pr[R \mid \hat{W}] = \Pr[R \mid W^*]$, and by (i) the latter is A^*. So the second term of Eq. (6) is zero and $A^* - \hat{A} = (1 - \hat{p})(A^* - \hat{c}_0)$. If $A^* \neq \hat{c}_0$ this gives $1 - \hat{p} = (A^* - \hat{A})/(A^* - \hat{c}_0)$; if $A^* = \hat{c}_0$ then $\hat{A} = A^*$ and the write failure rate is not identified from accuracies, because resolving the locator is worth nothing to this read loop.

(iv) Assume $A^* > \sigma(T) \geq \hat{A}$ and $\hat{c}_0 \leq \sigma(T)$. First, route (iii) of Proposition 2 for the read loop ρ requires that ρ's accuracy on resolved ledgers be at most $\sigma(T)$; on the oracle ledger that accuracy is $A^* > \sigma(T)$ by (i), so the read loop does clear the measured bound when the ledger resolves the locator, and the read-side route is excluded for ρ on this family. Second, $\hat{c}_0 \leq \sigma(T) < A^*$ gives $A^* - \hat{c}_0 > 0$. From $\hat{A} \leq \sigma(T)$ and Eq. (6),

$$A^* - \sigma(T) \;\leq\; A^* - \hat{A} \;=\; (1 - \hat{p})\,(A^* - \hat{c}_0) + \hat{p}\,(A^* - \hat{c}_1),$$

so $(1 - \hat{p})(A^* - \hat{c}_0) \geq A^* - \sigma(T) + \hat{p}\,(\hat{c}_1 - A^*)$, and dividing by $A^* - \hat{c}_0 > 0$ gives the general bound. Under Assumption 2 the term $\hat{p}(\hat{c}_1 - A^*)$ vanishes and the bound becomes $1 - \hat{p} \geq$

$(A^* - \sigma(T))/(A^* - \hat{c}_0)$, which is strictly positive since the numerator is. Hence the model composite fails to resolve a positive fraction of locators, and the loss $A^* - \hat{A}$ is attributed to the write in the amount Eq. (6) states. □

Remark on measurability. Every quantity in Eq. (6) is computable from artefacts the grid already produces. $\hat{W}$ is decided per item by comparing the model ledger's resolver output with the annotated locator; $\hat{c}_1$ and $\hat{c}_0$ are the model composite's accuracies on the two resulting item subsets; A^* is one further arm over the same items. The second term of Eq. (6) is therefore not a nuisance parameter to be assumed away but a number, and Assumption 2 is a hypothesis about that number being zero which the same artefacts test. The oracle write itself uses the annotations, which the model composite never sees; that is what makes A^* an upper reference for this read loop rather than an achievable arm.

PROPOSITION 6 (THE LOCATOR IS DEFINED ONLY RELATIVE TO THE RELATION)

Proof. Fix a stream $a_{1:N}$, an event type E, and relations $\sim_1 \subseteq \sim_2$, meaning that $u \sim_1 v$ implies $u \sim_2 v$ for all $u, v \in \mathcal{S}$. Write Occ_i, m_i and $L_i(k)$ for $\mathrm{Occ}_{\sim_i}(E)$, $m_{\sim_i}(E)$ and $L_{\sim_i}(E,k)$.

(i) If $t \in \mathrm{Occ}_1$ then $a_t \sim_1 E$, hence $a_t \sim_2 E$, hence $t \in \mathrm{Occ}_2$; so $\mathrm{Occ}_1 \subseteq \mathrm{Occ}_2$ and $m_1 \le m_2$. If $L_1(k)$ is defined then $k \le m_1 \le m_2$, so $L_2(k)$ is defined. The k smallest elements of Occ_1 are k distinct elements of Occ_2, each at most $L_1(k)$; so Occ_2 contains at least k elements that are at most $L_1(k)$, and its k-th smallest element satisfies $L_2(k) \le L_1(k)$.

(ii) Let $D = \{t < L_1(k) : t \in \mathrm{Occ}_2 \setminus \mathrm{Occ}_1\}$. The elements of Occ_2 strictly below $L_1(k)$ are the elements of Occ_1 strictly below $L_1(k)$, of which there are exactly $k-1$, together with the elements of D. If $D = \varnothing$, then Occ_2 has exactly $k-1$ elements below $L_1(k)$ and, by (i), contains $L_1(k)$ itself; so $L_2(k) = L_1(k)$. If $D \neq \varnothing$, then Occ_2 has at least k elements strictly below $L_1(k)$, so its k-th smallest element is strictly below $L_1(k)$ and $L_2(k) < L_1(k)$. Hence $L_2(k) = L_1(k)$ if and only if $D = \varnothing$.

(iii) A RET item's premise under $\sim_i$ is $m_i = 1$; being well formed under $\sim_1$ means $m_1 = 1$, which by (i) allows any $m_2 \ge 1$, so the item is well formed under $\sim_2$ exactly when $m_2 = 1$. A TRK item's gold under $\sim_i$ is m_i; since $\mathrm{Occ}_1 \subseteq \mathrm{Occ}_2$, $m_2 = m_1$ if and only if the two sets are equal, so the gold changes exactly when they differ. A CONJ item's locator under $\sim_i$ is $L_i(k)$, so the locator moves exactly when $L_2(k) \neq L_1(k)$, which by (ii) is exactly when $D \neq \varnothing$. Admissibility of the option set under $\sim_2$ requires its members to be pairwise $\sim_2$-inequivalent (Definition 4); pairwise $\sim_1$-inequivalence does not imply this, since $\sim_2$ is coarser, so the option set remains admissible only if no two options are $\sim_2$-equivalent.

(iv) The set of well-formed items under $\sim$ is determined by the predicates in (iii), the gold of each item by $m_\sim$ or by $V(x_{L_\sim(E,k)+1})$, and both depend on $\sim$ through $\mathrm{Occ}_\sim$. The ceiling of Proposition 3 on an item is $1/m_\sim(E)$, and the ceiling of Proposition 4 is a function of the value position $t = L_\sim(E,k)+1$ through the event $\{t > N - L\}$; both therefore change with $\sim$ whenever (i)–(ii) change m or L. The ceiling sum $\bar{\sigma}(T)$ is an item average of these over a family whose membership itself depends on $\sim$, so it has no value until $\sim$ is fixed. □

Remark. The inclusion $\sim_{\mathrm{raw}} \subseteq \sim_{\mathrm{cls}}$ is what makes the direction in (i) uniform: passing from strings to classes never removes an occurrence and never postpones the k-th one. Amendment A0-A3 of the registered record is the instance: under the raw relation the generator counted spellings, and every one of the three predicates in (iii) was false on a substantial share of items when evaluated under the class relation. The corrected instrument fixes $\sim = \sim_{\mathrm{cls}}$ and adds the $\sim_{\mathrm{cls}}$-inequivalence gate on options, which is the admissibility condition of (iii).

LEMMA 1 (A CLASS-LEVEL PERCEIVER CANNOT SEPARATE CLASS-EQUIVALENT OPTIONS)

Proof. Let the perceiver have located the value frame x_t exactly and read its class correctly, so that it knows $\pi_\sim(a)$, the class of the gold. Everything it reads from any frame factors through $\pi_\sim$, so its observation determines, about the gold, exactly that $a \in O_a$: every $o \in O_a$ is $\sim$-equivalent to a, hence has the same class, hence is consistent with every class-level observation the perceiver can make on this or any other frame; and every $o \notin O_a$ is excluded by the class it read. Beyond q and

O the perceiver's information therefore narrows the gold to O_a and nothing finer, and by the gates its success probability is at most $1/|O_a|$. If $|O_a| \geq 2$ this is strictly below 1, so no $\sim$-perceiver answers the item with certainty. The rule that returns a uniformly random element of O_a succeeds with probability exactly $1/|O_a|$, so the bound is attained. Since the location step and the class step were granted, the bound holds a fortiori for a perceiver that must also locate the frame and read it. □

Remark. The lemma is the reason the option gate in Definition 4 is stated in terms of the relation rather than of string distinctness. A single-frame write path whose vocabulary is the class vocabulary of the corpus is a $\sim_{\mathrm{cls}}$-perceiver; on an item generated under $\sim_{\mathrm{raw}}$ with two class-equivalent options it is bounded away from the gold by the item and not by its access structure, and a null scored against such items cannot be read as a null about composition. Under $\sim_{\mathrm{cls}}$ with the gate, $|O_a| = 1$ on every item and the lemma imposes nothing, which is the state the corrected instrument is in.

O REPRODUCTION

Annotations are fetched from the corpora's public servers; no media is redistributed. Keyframes are extracted from EPIC-KITCHENS over HTTP byte ranges, so reproducing the corpus requires no bulk download and no signed agreement. The programs are, in the order the pipeline runs them: `a0_conjfrac_probe.py` (the incidence instrument), `a0_corpus.py` (the frozen generator; `--occ-key class --ret-pool verb_matched` for v3), `a0_bounds.py` (the ceilings of Table 1), `epic_score.py` (the judge-free scorer), `run_a0.py` and `run_a0_v3.sh` (the grids, one directory per corpus version), the companion's `run_axis.py` with `AXIS_DATA=data/epic_v3` for the oracle and model-ledger arms of Table 16, and `a0_tables.py` / `a0_v3_tables.py` (every table and macro in this paper, written from the per-question records), `axis_decomp.py` (Table 9) and `a0_stats.py` (the intervals and paired tests of Tables 8 and 13, computed from the same records and from nothing else). Those records are released for every cell, which is what makes the tests checkable rather than assertable. Every cell this paper reports has landed; no number in it is a placeholder.

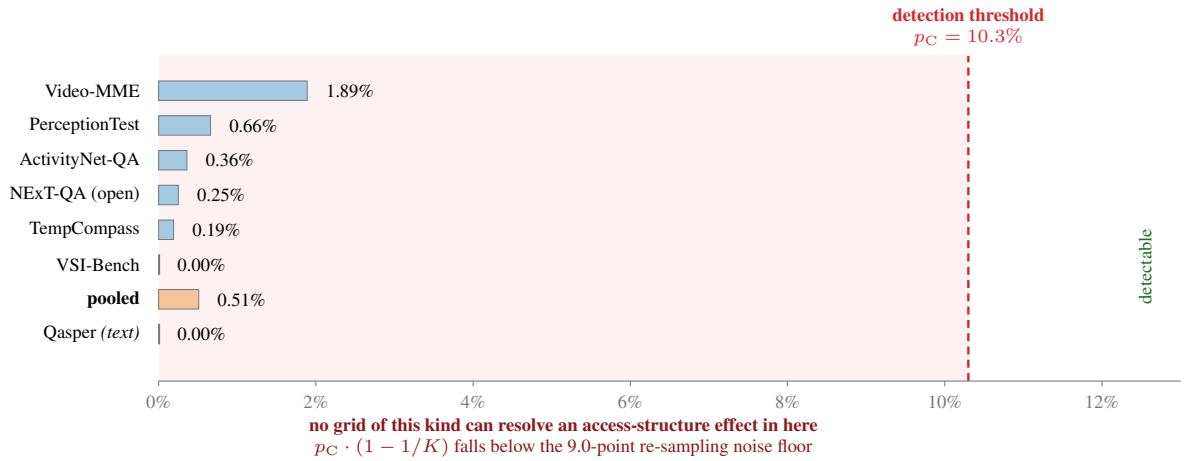


Figure 13: The conjunctive fraction, against the fraction a grid of this kind would need. Bars are the measured lower bound of Table 2; the shaded band is where an access-structure effect cannot be resolved at all, because even a composite that solved every conjunctive item perfectly and paid nothing elsewhere would move a benchmark mean by $p_C \cdot (1 - 1/K)$, which stays under the 9.0-point re-sampling noise floor a companion measures for exactly this grid (Anonymous authors, 2026d). Every corpus sits inside the band (even the richest, Video-MME at 1.89 %, by more than a factor of five), and the pooled figure by 20×. VSI-Bench, one of the three natural-video benchmarks a companion registered, contains no conjunctive query at all; the text control contains none either, which is the asymmetry CCH's Limitation (iii) predicts. The claim this figure supports is about instruments, not architectures, and one measurement falsifies it: a public corpus with $p_C > 10.3\%$.

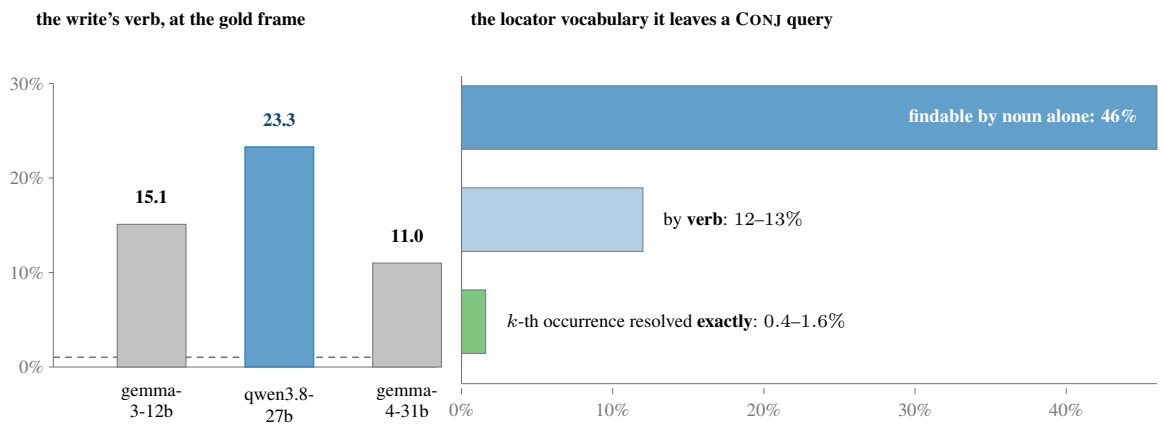


Figure 14: Upstream of every arm: what the shared single-frame write can decide. Left: verb correct at the annotated segment midpoint, the most favourable single frame, scored through the corpus's published class tables (§6.1). The dashed line marks the uniform rate over the ∼97 verb classes, and the bars are the identity factor 0.11–0.23 that every arm inherits. The errors are concentrated on time-reversal pairs (gold *put* (n=128) is written *take* 46% of the time by the most verb-accurate backbone, gold *close* (n=46) written *open* 26%), which is Theorem 1 of the companion (Anonymous authors, 2026e) read off this cache. **Right:** the vocabulary a CONJ locator ("the k-th time you *verb noun*") is left with under that write. A retriever cannot address the k-th occurrence of an event the index does not name, and a counter cannot count it: the CONJ column of Figure 3 is this factor's product with the ordinal factor, measured from the read side.

P ADDITIONAL FIGURES AND TABLES

Every figure and table of the full manuscript that does not appear in the main text is collected here, in the order the argument uses them: the conjunctive fraction against the fraction a grid of this kind would need (Fig. 13), the shared single-frame write path (Fig. 14), the construction of a conjunctive item (Fig. 15), the witness corpus and the witness-validity ledger (Tables 15 and 16) the reversal of the dual ordering (Fig. 16) and the witness contrast (Fig. 17).

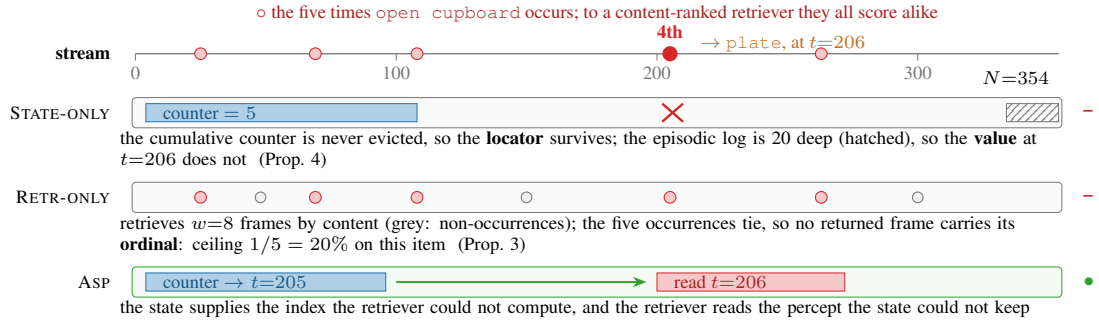


Figure 15: Why the conjunctive family needs both channels, on one real item (`P01_14`: *"Consider the fourth time you open the cupboard . . . what did you interact with immediately after that?"*). `open cupboard` occurs five times, and the values following the other four (`cup`, `board:cutting`, `colander`, `glass`) are four of the item's eight options, so landing on the wrong occurrence scores zero by construction rather than by luck. Neither channel alone reaches the answer, and they fail for different reasons, which is what makes the family conjunctive rather than merely hard. Averaged over the family the index ceiling is 31.2% and the state ceiling 23.6% (Table 1); this item sits at $1/5$ on the first.

Table 15: The witness corpus. EPIC-KITCHENS-100 validation split, the 8 longest streams by action count. One keyframe per annotated action segment; 3,092 keyframes, 735 items balanced across the three families. Gold is an option letter derived from held-out annotation, so no judge is involved.

stream	P01_14	P01_15	P02_12	P22_01	P22_03	P24_09	P29_05	P30_08	total
keyframes N	354	291	371	350	484	347	564	331	3,092

Table 16: Witness validity: the same read over a validated ledger. gemma-3-12b, v3 corpus, B=4096; measured additive bound on CONJ is 12.3. Middle block: the address-structured read over the ledger the model's own write produced. Lower block: the *same* read loop over a ledger built from the corpus's annotations, declared an oracle in every use. Superscripts as in Table 4.

arm	RET	TRK	CONJ
BLIND	18.5	5.2	17.2
STATE-ONLY	22.4[7]	12.0[2]	16.9[5]
RETR-ONLY	10.7	22.4	12.7
ASP	14.9[1]	16.6	16.2
AXIS (model ledger)	14.3	10.4[11]	18.8
AXIS, wrong address	16.6	10.7[12]	18.5
AXIS, **oracle ledger**	**20.8**	**73.0**[1]	**31.5**
oracle, no dereference	13.6	72.4[1]	13.6
oracle, wrong address	20.1	74.0[1]	17.2

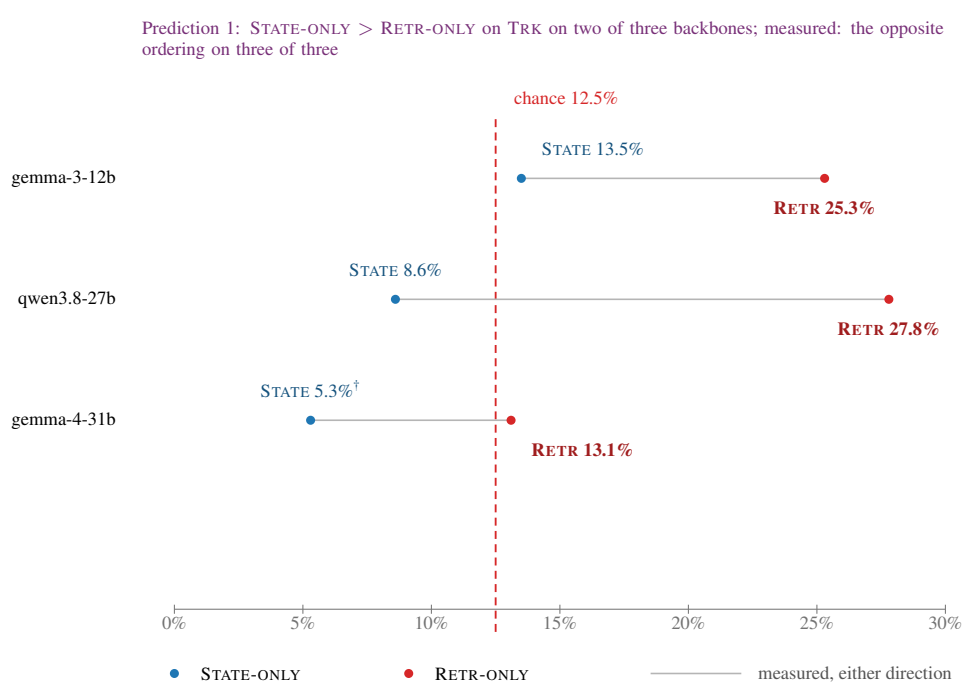


Figure 16: The dual ordering the corpus was built to test, and its reversal. Prediction 1 expected the state channel to win the counting family, whose ledger holds the counter; the measured grid has RETR-ONLY ahead on every backbone (Table 4): $25.3/27.8/13.1$ against $13.5/8.6/5.3^{\dagger}$. The counting mechanism is in §E.1: the captions carry the verb–noun pair the counter needs, while the state's `action_counts` over-count a persisting action (216 counted on a 354-step stream). gemma-4-31b's STATE-ONLY cell is marked †: 137 of 245 predictions carry no option letter, so the cell is reported and not scored (§6), and the reversal is established on the two scorable backbones. Drawn from the same generator output as Table 4.

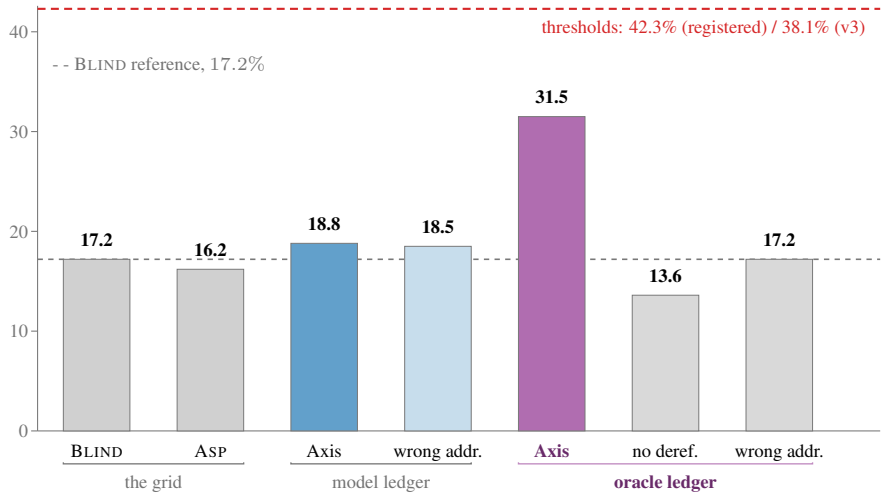


Figure 17: The witness is valid: the oracle separates, and the prediction still fails. Amendment A0-A2's test (A0-W1-v3), Table 16 in bar form: CONJ accuracy (%), gemma-3-12b, v3, B=4096; same read loop and dereference throughout. Bars, left to right: BLIND and ASP from the grid; the companion's address-structured read (AXIS) over the model ledger and over its wrong-address control; and over the corpus's annotation (oracle) ledger with its no-dereference and wrong-address controls. The validity criterion holds: on CONJ the oracle clears both controls by $+17.9$ and $+14.3$ points while on TRK it separates from neither (73.0 vs $72.4/74.0$: the answer is the ledger's count, the frame carries nothing). Against the registered threshold 42.3% and the v3 threshold 38.1%, prediction A0-W1-v3 **fails**; on gemma-4-31b the oracle cell scores 37.7% (9 of 308 unparsed, scorable) and also fails, and on qwen3.8-27b it scores 29.2% with 19 unparsed, above this paper's own 5% line: reported, not scored. The loss is on the write side, and the next factor is the read of one addressed frame (§6.2).