What language buys: re-sampling, transmission, and the limits of a discriminability metric

W. Schulz

Jul 30, 2026

Abstract

Eight run-once experiments split language into measurable layers. Extra reasoning steps improve discrimination only when the input can be re-sampled; the metric is provably blind to compositionality; efficiency fixes a lexicon's shape but not its identity. Several hypotheses were withdrawn and two experiments voided as design failures.

Eight experiments and one paired re-test, run once each on 29–30 July 2026, dissecting language into layers a machine can measure. The nine scores were 2/7, 1/4, 4/4, 4/5, 3/4, 2/3, 5/5, 1/4 and 0/1, and — as usual in this project — the informative results were mostly the misses. Two hypotheses were withdrawn outright, one mechanism turned out to be mostly wrong under a paired re-test, two experiments were voided as design failures, a later prior-art sweep took back three more claims, and one result is a two-line proof that closes off an entire line of attack.

Depth buys discrimination, but only from evidence you can re-sample

The campaign opened on a slogan worth testing: that no amount of thinking repairs a distinction destroyed at encode time.

Take one recurrent architecture, one task, one noise scale, matched training depth, and vary only what the *input* does. When the encoding is re-drawn every step, node accuracy goes from 0.978 at sixteen steps of thinking to a perfect 1.000 at 256, with discriminability d' growing at a log-log slope of 0.519 — the $\sqrt{k}$ rate, which is Wald's sequential-probability-ratio result from 1945 arriving as a control that works. When the encoding is drawn once and frozen, accuracy is 0.935 at sixteen steps and 0.935 at 256. Flat.

The mechanism is not inferred; it is measured in the increments. Under the frozen encoding, the per-step change in the hidden state decays by a factor of 347 — the contractive update converges to a fixed point, and that fixed point contains exactly the information in the one look it got. Under the re-sampled encoding the same network never settles, its increment holding steady, and it accumulates to perfection.

> No amount of thinking repairs a distinction destroyed at encode time; only re-encoding does.

A ceiling rather than a wall — the frozen network still reaches 0.935 — but the ceiling is flat from thirty-two steps to 256 while the re-samplable arm climbs to 1.000. And the half of the original framing that is simply *false* is worth stating: "test-time compute buys composition, not discrimination" is wrong. Depth buys discrimination fine. What it cannot buy is discrimination from a sample it cannot take again.

One retraction, visible. A companion experiment had proposed that contractivity *caps* evidence integration — that the gate this project introduced to stop networks from overthinking also prevents them from accumulating, making stability and integration one knob pointed in opposite directions. The preregistration named the condition under which that must be withdrawn, and the condition occurred. The hypothesis was backwards. The non-contractive residual network does not integrate; it *degrades* with depth under noise, and collapses outright when trained long. The contractive network is the one that integrates. The binding constraint was the training-depth distribution all along.

The correct statement is a correction to how this repo had been describing contraction: contractivity is defined relative to a fixed input. Under a re-sampled input the same map is a stable tracker, not a converger. It is a property of the map *and its input stream*, not of the map alone.

And one result that transfers to a shipped tool. A converged consistency residual is a *halt* signal, not a *correctness* signal. Measured on the frozen arm, the residual predicts error at an AUROC of 0.589, and 82% of wrong answers are fully self-consistent — residual exactly zero. The adaptive-halting tool this project ships still stands, since the residual beats the best fixed depth on ten of ten seeds; but "settled" carries almost no information about "correct." The general form:

> Internal consistency cannot detect a naming error. A system given the wrong word builds a flawlessly coherent, entirely wrong world, and every check it can run on itself will pass.

Compositionality is invisible to the instrument this project uses

The first two experiments measured encoding fidelity, which is a statement about *distinctions* wearing a language costume — nothing in them required a lexicon, a symbol, or a second party. The third tested what actually makes a language a language: a finite reusable vocabulary, composition, and transmission to a second learner.

Every code tested is a bijection from 256 meanings onto the same message space, so capacity is matched *exactly*, and the codes differ only in how meanings map to messages. The result is a total dissociation across three payoffs.

Against noise, the compositional code and the holistic code are identical: error rates of 0.380 and 0.375. Against the shape of errors, they diverge completely: given an error, the compositional code lands 1.18 attributes away and the holistic code 3.04, where chance is 3.00. Against a new learner shown only part of the meaning space, they diverge totally again: 1.000 against 0.250, where chance is 0.250.

> Compositionality buys nothing against noise and everything against a new learner.

And the middle claim is not an experiment, it is a proof. Under position-wise embedding of a shared vocabulary, every one of these codes is a bijection onto the same set of points in embedding space — verified numerically at a maximum difference of 0.000e+00 after sorting. The codes differ only in which meaning is attached to which point, and a nearest-codeword decoder's error rate cannot see that. Therefore:

> Exploitability — the discriminability quantity this program has measured across four campaigns — is provably blind to compositionality. At matched capacity, the property that makes language language is exactly invisible to the instrument the program has been using.

That is why a distinction-based framework could not see language. Language's payoff is not in the signal-to-noise of one agent's cuts; it is in what a *second* agent can do with them, and in where the code

fails when it fails.

One pre-registration error is reported as one rather than quietly fixed. The tolerance set for the error-rate spread was 0.005, and the observed spread was 0.00514 — a formal miss. But the Monte Carlo standard error of a single estimate in that design is 0.00215, so the expected maximum spread across the grid is about 0.0075. The threshold was tighter than the sampling noise, and should have been computed in advance. The claim itself is not in doubt, because the point-set check returns exactly zero. The identity holds; the numerical test of it was invalid.

Speech and writing are not two notations for one thing

If re-samplability is the axis that matters, then writing is not a way of recording speech. It is the technology that converts an un-re-samplable channel into a re-samplable one — which means an earlier experiment about frozen encodings had, unknowingly, been an experiment about literacy.

Four conditions with total exposure held equal at three looks per symbol: one hearing; blocked repetition, each symbol said three times in a row; interleaved re-passes, the whole message three times through; and a page, with all looks available in parallel and no working memory at all in the decoder.

The persistent page beats the best transient condition by +0.096 when memory is scarce relative to message length, and by +0.001 when memory is ample. The interaction is clean:

> Writing buys nothing when the listener has enough working memory, and buys a great deal when they do not — and the shortfall grows with message length. The advantage of the written form is not fidelity, resolution, or capacity; those are matched here by construction. It is that the page holds the partial results, so the reader does not have to.

That also explains what long, syntactically deep sentences have to do with literacy. At short message lengths the ear is already nearly at the page; at double the length it falls behind by a tenth. Length is the variable that converts a memory bound into a comprehension bound.

The blocked-versus-interleaved prediction came out in the right direction but underpowered, so it was re-tested with a paired analysis — same code, same cell, changing only the analysis and the sample, disclosed as a power fix rather than a second bite. Pairing worked as an instrument: the standard error fell from ±0.05 to ±0.0041, a twelvefold power gain, confirming that the variance had been entirely the shared symbol-embedding draw.

The effect is real and the explanation for it was mostly wrong. Blocked repetition does beat interleaved re-passes at matched exposure — fifteen of sixteen seeds positive across both cells, with mean-over-standard-error of 4.0 and 6.7. But it is small, +0.017 against a predicted 0.05, and at eight times the working memory the effect shrinks by only about 43% while becoming *more* reliable. If a working-memory bound were the mechanism, ample memory should have abolished it. So working memory accounts for at most ~43%, and the remainder is most likely temporal distance between related evidence — in blocked order the three looks at a symbol are two steps apart, in interleaved order they span the whole message, which is harder for credit assignment through a recurrence regardless of capacity. That is a *learnability* effect, not a *capacity* effect, and the pre-registration did not distinguish them.

This costs the mapping onto oral-formulaic composition something real. Both readings are relevant to a singer working without a text, but only the capacity reading is what the classic claim asserts, and the capacity reading is not supported here. A clean discriminator exists and is cheap: hold memory fixed and vary only the gap between a symbol's repeated looks. It also makes the page-versus-transient result

more trustworthy by comparison, since that one does collapse to zero with ample memory exactly as a capacity account requires.

The falsifier: efficiency explains the shape of a lexicon, not its identity

The synthesis emerging from all of this — that every payoff of language lives *across* a boundary — was too loose to be worth anything, so it was sharpened before being tested, and the sharpening cost two of the three examples it had been built on:

> A payoff is a crossing payoff if and only if it requires a decoder that does not already hold the code.

Under that definition, more looks and memory-on-the-page are *not* crossings; they are bounded-resource effects inside one agent. Only transmission survives. The defensible residue: exploitability presupposes that the decoder already holds the codebook. Language is what happens when it does not.

Layer six — economy — was then chosen as the falsifier, because its payoff looks like it belongs to a single agent at a single moment: a good lexicon carves a meaning space efficiently, and you can evaluate that alone with the codebook in hand. Meanings are drawn from a six-cluster mixture; a lexicon is a set of centroids; rate is matched exactly; eight independently initialised optimal lexicons are built per vocabulary size, and cross-understanding is scored with the *exact* best index alignment via a solver validated against brute force.

Two equally efficient eight-word lexicons understand each other 2.1× worse than themselves, even under optimal alignment. At sixteen words, 3.1× worse. Degeneracy grows monotonically with vocabulary size.

> Efficiency explains the *shape* of a lexicon; it cannot explain its *identity*.

A first reading of the table also claimed that efficiency *uniquely determines* a small lexicon — the four-word row showed all eight optimisations finding the same partition. That was flagged as single-world and needing a check; the check was run across five worlds and it failed. Four-word lexicons were unique in only two of five worlds, and in one world a four-word lexicon is already 4.2× degenerate. The claim is withdrawn.

A replacement mechanism was proposed on the spot — that degeneracy tracks the mismatch between vocabulary size and the number of natural kinds in the world, rather than vocabulary size as such, so the degeneracy curve should be U-shaped with its minimum where the vocabulary matches the world's joints. Two attempts to test that both failed as instruments, and it remains untested. The story is recorded here as an unsupported hypothesis, not a finding, because the way it failed is more useful than the story was.

Read against the literature, what *does* survive is a precise correction to how the efficiency account of vocabulary should be taken. Real lexicons being near-optimal is necessary and nowhere near sufficient to specify one. The efficiency account explains why lexicons look the way they do without explaining why a community has *this* one — and the residue is a coordination problem, which these numbers locate and measure.

The most persuasive result in the campaign was the invalid one

The mismatch test was fully pre-registered, scored 1/4, and both of its kill conditions fired — the most convincing output an experiment can produce. Every verdict was void.

It inherited the previous experiment's code, its metric and its exactly-validated alignment solver, but not its validity gate: the requirement that the lexicons being compared are genuinely *equally efficient*, a distortion spread of at most 0.15, which the parent passed at 0.083. The child never checked. Measured afterwards, its spreads were 2.16 to 2.46 — best-quality lexicon 0.60 against worst 13.34, a 22× gap. It had been comparing good lexicons against broken ones and calling the difference degeneracy. What it actually measured was the failure rate of the clustering algorithm's random initialisation.

The irony is worth keeping. The design used *well-separated* clusters precisely to make "natural kinds" unambiguous — and well-separated clusters make random initialisation worse, because an initialisation that drops two centroids inside one cluster and none in another is catastrophically costly and the algorithm cannot escape it.

A repair attempt enforced the premise directly: forty initialisations per cell, keep the best eight, and refuse to report any cell whose survivors still were not equally efficient. The gate failed anyway, four cells of nine, so the repair declared a design failure and voided its own remaining predictions rather than reporting them. That is the behaviour the method is supposed to produce, and it leaves the question simply unanswered.

A fourth explanation is available and was deliberately not pursued — that degeneracy may require *overlapping* structure, since a world with sharp joints has a genuinely best carving while a smooth world does not. It is plausible, it fits both the surviving result and the failure, and it is exactly the kind of fluent post-hoc story this project's gate exists to distrust. Three post-hoc explanations had already been killed in a row. It is recorded for a future pre-registration rather than tested as a rescue.

> When you extend an experiment, carry its controls forward. A pre-registered kill condition firing is evidence only if the measurement's premises were *checked*, not assumed.

What a late prior-art sweep took back

The sweep assigned to this territory had died partway through and was resumed after the campaign closed. It cost three corrections, which are applied here rather than left in the working notes.

The transmission row is near-analytic too. The three payoffs were presented as one provable and two measured. In fact two of the three are forced. A holistic code is a *random* bijection, so for a meaning the learner has never seen there is, by construction, zero information anywhere in its training data. The 0.250 is chance because chance is the only available answer. The genuinely empirical rows are the error-shape result and the graded ladder between them. This also fixes the relation to the literature: the published claim that compositionality is *not* necessary for generalisation concerns emergent codes, where structure is discovered rather than imposed. A random bijection is a strictly harder and degenerate case, so these experiments are not evidence about that claim and must not be cited as though they were.

The motivating example was false. The whole line was inspired by tokenisation — the familiar failure where a model miscounts the letters in a word. That intuition does not survive contact with the literature: chain-of-thought approaches *recover* tokenisation-induced deficits (shown for arithmetic under different number tokenisations), and the gap shrinks with scale; subword embeddings robustly encode character-level information, recoverable by a probe, across several scripts. The distinctions are present and elicitable, not destroyed. It is an elicitation failure, not an encoding failure. The frozen-encoding result here genuinely destroys information — a noise draw that cannot be re-taken — which is a far stronger condition, and the result must not be carried over to tokenisers.

And the general form of the slogan is not merely empty but false. "Depth cannot buy discriminability" is the data-processing inequality for Shannon information, which is empty; in the *usable*-information sense it is wrong, since predictive information can be created through computation in explicit violation of that inequality. What survives is narrow and architectural: *this* bounded network, on *this* task, cannot extract more from a sample it cannot re-take.

What the sweep cleared: the compositionality metric used here survives its own critical literature, because those critiques attack *inferring* structure from a metric value in a code the experimenter did not control. Here compositionality is an independent variable set by construction and the metric is a manipulation check on a known quantity — a legitimate use, provided it is never called a measurement.

What is not claimed

The $\sqrt{k}$ integration result is Wald's. The zero-margin limit is the data-processing inequality and is empty. That a settled network stops learning from a fixed input is Banach's fixed-point theorem. The transmission result is Kirby's and Chaabouni's, the coding-theory direction is Shannon's and Hamming's, the orality material is Parry's, Lord's, and Ong's, the memory model is Baddeley and Hitch's, and the efficiency baseline is Lloyd–Max. The complexity classes reachable by more thinking steps are a published theorem, and so is the taxonomy of where extra thinking helps at scale. Nothing in the campaign is filable or publishable.

What is contingent, measured, and this project's own is the set of numbers separating the regimes — and the sharpest of them, the invariance identity, is two lines of linear algebra that nonetheless close a question an adversarial sweep had independently flagged as the highest-value open one in the territory. That convergence is worth something on its own: the sweep, working from the literature and blind to these experiments, named "is the efficient-lexicon optimum unique or degenerate?" as the single unchecked residue with the right shape — which is exactly the question the falsifier answered.

Sources

The numbers in this chapter are this project's own run-once results (`language/RESULTS-L1.txt` … `RESULTS-L6b.txt` and `language/FINDINGS.md` in the Quantum-Intelligence repository). The works below are the prior art the chapter names.

1. Wald, A. (1945). "Sequential Tests of Statistical Hypotheses." *Annals of Mathematical Statistics* 16(2): 117–186. doi:10.1214/aoms/1177731118
2. Banach, S. (1922). "Sur les opérations dans les ensembles abstraits et leur application aux équations intégrales." *Fundamenta Mathematicae* 3: 133–181. doi:10.4064/fm-3-1-133-181
3. Shannon, C. E. (1948). "A Mathematical Theory of Communication." *Bell System Technical Journal* 27(3): 379–423. doi:10.1002/j.1538-7305.1948.tb01338.x
4. Hamming, R. W. (1950). "Error Detecting and Error Correcting Codes." *Bell System Technical Journal* 29(2): 147–160. doi:10.1002/j.1538-7305.1950.tb00463.x
5. Kirby, S., Cornish, H., & Smith, K. (2008). "Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language." *PNAS* 105(31): 10681–10686. doi:10.1073/pnas.0707835105
6. Chaabouni, R., Kharitonov, E., Bouchacourt, D., Dupoux, E., & Baroni, M. (2020). "Compositionality and Generalization in Emergent Languages." *ACL 2020*. arXiv:2004.09124 — the claim about *emergent* codes that the transmission row must not be cited against.

7. Singh, A. K. & Strouse, D. J. (2024). "Tokenization counts: the impact of tokenization on arithmetic in frontier LLMs." arXiv:2402.14903
8. Kaushal, A. & Mahowald, K. (2022). "What do tokens know about their characters and how do they know it?" *NAACL 2022*. arXiv:2206.02608
9. Xu, Y., Zhao, S., Song, J., Stewart, R., & Ermon, S. (2020). "A Theory of Usable Information Under Computational Constraints." *ICLR 2020*. arXiv:2002.10689 — V-information "can be created through computation," in violation of the data-processing inequality.
10. Merrill, W. & Sabharwal, A. (2024). "The Expressive Power of Transformers with Chain of Thought." *ICLR 2024*. arXiv:2310.07923
11. Sprague, Z., Yin, F., Rodriguez, J. D., et al. (2024). "To CoT or not to CoT? Chain-of-thought helps mainly on math and symbolic reasoning." arXiv:2409.12183
12. Baddeley, A. D. & Hitch, G. (1974). "Working Memory." *Psychology of Learning and Motivation* 8: 47–89. doi:10.1016/S0079-7421(08)60452-1
13. Lloyd, S. P. (1982). "Least squares quantization in PCM." *IEEE Transactions on Information Theory* 28(2): 129–137. doi:10.1109/TIT.1982.1056489; Max, J. (1960). "Quantizing for minimum distortion." *IRE Transactions on Information Theory* 6(1): 7–12. doi:10.1109/TIT.1960.1057548
14. Parry, M. (1971). *The Making of Homeric Verse: The Collected Papers of Milman Parry*; Lord, A. B. (1960). *The Singer of Tales*; Ong, W. J. (1982). *Orality and Literacy*.

Published by NeuroGenesys (AW3 Technology, Inc.). Canonical version: https://neurogenesys.si/research/what-language-buys