The Compression Fallacy

Why Shorter Is Not the Same as Better Understood

Flyxion

July 2026

Abstract

A widely held assumption, rarely stated but frequently acted on, treats the shortest description of a pattern as the correct measure of how well that pattern is understood: a good model is a compressed model, and compression is the currency of understanding. This paper argues that the assumption outruns what has actually been established about the relationship between description length and understanding. Using a formal apparatus already developed for measuring representational capacity—an ontological triple of a domain, an indistinguishability relation, and a generating ontology, together with two independently defined deficits, one for excess coding length and one for lost distinguishing capacity—the paper shows that the proven relationship between the two runs in only one direction: perfect compression guarantees perfect distinguishing capacity, under an assumption about faithful encoding that need not hold for real representational systems, but the converse remains an open conjecture. Treating description length as a continuous, graded proxy for understanding, as is now common practice in discussions of machine-learned representations, extends this one-directional and conditional result well past what it supports. The Forth programming language supplies a worked case in which the extremes come apart cleanly: Forth code is short not because it approaches the Kolmogorov-optimal description of the task it solves, but because it is reachable, at low cost, from an already-accumulated library of admissible operations—a distinct and prior achievement that the brevity of the resulting code does not, by itself, report. The paper closes by proposing that a model's distinguishing capacity, not its description length, is the quantity that should be measured directly wherever the two can be pulled apart.

Contents

1 An Old Definition, Asked of a New Case

The computer scientist Daniel Hillis once offered a definition of the information content of a pattern: the amount of information in a pattern of bits equals the length of the smallest computer program capable of generating those bits. The definition is, in substance, Kolmogorov complexity restated for a general audience, and Hillis's own statement of it claims nothing beyond that restatement. What this paper is concerned with is a further claim the definition has often been taken, by others, to support, though it is not one Hillis himself is on record as drawing: that a system which produces a shorter description of something has thereby understood that thing better than a system producing a longer one. The claim is rarely defended directly. It is simply assumed, in the background of arguments about which model of a phenomenon is the better model, which representation of a domain is the more insightful representation, and—increasingly, in discussions of large trained models—which network has learned the most genuine structure rather than having merely memorized surface statistics.

Forth, an unusually terse programming language whose devotees have long pointed to Hillis's definition as evidence of the language's virtue, supplies an unusually clean test case for the assumption, because Forth's brevity is well documented, uncontroversial, and produced by a mechanism that is easy to state precisely: word-factored Forth code is short because a Forth programmer builds a library of named operations—words—and subsequent code invokes them rather than re-deriving them (Ziembik 2025). A task solved in Forth after such a library exists can be several times shorter, in raw character count, than the same task solved in a general-purpose language without access to an equivalent library. The question this paper asks is not whether Forth code is short; that is not in dispute. The

question is what that shortness is evidence of, and whether it is evidence of the thing Hillis's definition, read the way it is usually read, would have it be evidence of.

This paper argues that it is not, and that the same gap separating Forth's brevity from Kolmogorov-optimal description separates, more consequentially, model compactness from model understanding wherever the comparison is drawn. Section 2 introduces the formal apparatus needed to state the gap precisely, developed independently of any concern with Forth or with machine learning. Section 3 states the apparatus's central result exactly, including the asymmetry in what it does and does not establish, since the asymmetry is the paper's entire argument. Section 4 works through the Forth case as a concrete instance of the gap. Section 5 turns to the contemporary target: the informal but pervasive assumption, in discussions of trained models, that shorter or sparser representations are better understood ones. Section 6 argues that engineered, deliberately simplified learning environments and productive detours through greater complexity during problem-solving are further evidence against treating compression progress as monotonic with understanding. Section 7 proposes what should be measured instead, where the two quantities can be pulled apart and are not.

2 Coding Length and Distinguishing Capacity

A general framework for measuring representational capacity (Flyxion 2026)[1] takes as its basic object not a set of items to be described, nor a description language considered on its own, but the pairing of the two together with a notion of when two items count as the same for the purposes at hand. Formally, an *ontological triple* $(X, \sim, T)$ consists of a set X, an equivalence relation $\sim$ on X specifying which elements of X are treated as observationally indistinguishable, and an ontology T: a description language, or template hierarchy, determining which descriptions of elements of X are actually reachable, as opposed to merely existing in principle. The distinction between what a description language can express in principle and what it can reach in practice, given the templates and operations it has already accumulated, is the hinge the rest of this paper turns on.

Two deficits are defined over this structure, and they measure different things. The *coding deficit* of an object x, written $\delta_T(x)$, is the excess length of the shortest description of x reachable within T over the Kolmogorov-optimal description length of x—the length of the absolute shortest program capable of generating x, considered independently of any particular ontology. The *distinguishability deficit* of x, written $\Delta_T(x)$, is a different quantity entirely, built from two further pieces of notation worth naming on their own: let $D_T(x)$ denote the *distinguishing capacity* of T at x, the number of other objects that can be separated from x by descriptions reachable in T, and let $D^*(x)$ denote the maximum such capacity attainable by any ontology whatsoever. The distinguishability deficit is then $\Delta_T(x) = D^*(x) - D_T(x)$: the gap between what T actually distinguishes at x and what the best possible ontology could. The coding deficit asks how much longer a description is than it needs to be. The distinguishability deficit asks how much distinguishing power has been lost. Nothing in the definitions forces these two quantities to move together, and the relationship between them, rather than being assumed, is the object a further theorem is required to establish.

[1]This paper departs from the practice, followed elsewhere in this series, of not citing the author's own prior work, for the same reason given in a companion paper: the argument here depends on a specific prior formal result, and describing that result without attribution would make it impossible to check against its source.

3 What the Proxy Theorem Actually Proves

The relevant result, the Deficit Proxy Theorem, states that under an assumption of faithful encoding—that every distinction a description language can express costs at least one bit to express—the distinguishability deficit is bounded above by a constant multiple of the coding deficit: $\Delta_T(x) \leq C\,\delta_T(x)$. A direct consequence, stated in the source theorem itself, is that a coding deficit of zero forces a distinguishability deficit of zero: an ontology that achieves the Kolmogorov-optimal description of an object thereby also achieves the maximum possible distinguishing capacity for that object. This is the direction of the relationship that is actually proved.

The converse—that a distinguishability deficit of zero forces a coding deficit of zero, or more generally that reducing an already-small coding deficit further continues to track improvements in distinguishing capacity—is explicitly not established. The source theorem states it as a conjecture, resting on the further assumption that optimal coding uses every distinction available to it, an assumption with no independent argument offered in its favor and no obvious reason to expect it holds for representational systems built by processes other than deliberate, distinction-preserving design.

The asymmetry matters because the popular reading of Hillis's definition, and the informal practice built on readings like it, uses the coding deficit as a graded, continuous stand-in for the distinguishability deficit across its whole range, not merely at the single point where both happen to be zero. The Deficit Proxy Theorem licenses an inference from perfect compression to perfect distinguishing capacity. It does not license the inference, drawn constantly in informal discussion, that a further reduction in an already-imperfect coding deficit constitutes evidence of a further reduction in distinguishing capacity, still less that the amount of the one measures the amount of the other. An upper bound is not a two-sided estimate. A system can, consistently with everything the theorem proves, carry a large coding deficit while carrying almost no distinguishability deficit at all: the bound $\Delta_T(x) \leq C\,\delta_T(x)$ is satisfied trivially whenever $\Delta_T(x)$ is small, regardless of how large $\delta_T(x)$ happens to be. Nothing rules this case out. Section 4 argues that it is, in fact, close to the ordinary condition of a mature, well-factored code library.

4 Forth and Reachable Complexity

The source framework's own introduction, developed independently of this paper's concerns, names but does not develop a third notion relevant here: *reachable complexity*, the shortest description of an object expressible from a description language's current template hierarchy, which may be strictly longer than the Kolmogorov-optimal description—longer, that is, than the true value of $L^*(x)$ against which the coding deficit is measured. Forth supplies reachable complexity's first concrete instance. A Forth programmer accumulates, over the life of a project, a library of named words: short, composable operations built from more primitive ones, each capturing an operational distinction the programmer has found useful to keep separately addressable. Code written after such a library exists is short because it is reachable, cheaply, from that library—not because it approaches $L^*(x)$, the description length that would be achieved by an ontology with no library at all, starting from nothing, optimizing purely for brevity against the raw task.

This is not a minor qualification. It inverts the naive reading of Hillis's definition as applied to Forth. The naive reading takes Forth's brevity as evidence that Forth programs approach the information-theoretic floor of the tasks they solve, and hence as evidence that Forth, among languages, comes closest to expressing the true content of a computation.

The reachable-complexity reading takes the same brevity as evidence of something else: that a well-factored library of admissible operations was available to be reused. Two Forth programmers solving the same task with different accumulated word libraries will produce differently short code for it, and the difference in length reports which library was available, not which programmer better understood the task. A companion result, developed independently of any concern with Forth, states the point in general form: no valid compression mapping can be constructed without prior identification of the invariant structure of the space being compressed, and that identification is itself achieved only through a prior process of expansion—exploration, iteration, and the discovery of which structure is invariant in the first place (Flyxion 2026b). Forth's terseness is compression in exactly this downstream sense: it presupposes the expansion that built the word library, and reports that expansion's prior existence rather than reporting anything about the Kolmogorov-optimal floor of the task the short code happens to solve.

None of this counts against Forth, and it is worth being explicit that it does not, since the point of this section is not to relitigate a decades-old argument about programming-language design. A large, well-factored word library achieving low $\delta_T(x)$ relative to itself, while $L^*(x)$ remains a separate and largely inaccessible quantity, is exactly what a mature, high-distinguishing-capacity representational system should look like. The coding deficit measured relative to the library is doing real work in this case; what it cannot do, without the further, unproved direction of the Proxy Theorem, is stand in for a claim about how close the resulting code sits to the absolute informational floor of the problem, or, more importantly for what follows, for a claim about how much of the task's true distinguishing structure the library actually preserves.

5 The Same Move in Machine-Learned Representations

The informal assumption this paper has been arguing against is not confined to programming-language advocacy; it is closer to background theory in contemporary discussions of trained models, where a shorter, sparser, or more compressed internal representation is routinely treated as evidence of deeper or more genuine understanding, and description-length reduction has been a recognized modeling virtue since well before large neural networks existed, formalized in the minimum-description-length principle (Rissanen 1978). The assumption's most explicit and most influential statement extends compression progress well past a technical modeling virtue and treats it as the underlying currency of understanding itself: a unifying account of curiosity, creativity, and aesthetic interest across infants, mathematicians, artists, and artificial systems alike, in which subjective beauty and interestingness are identified with the rate at which an observer's compression of its data is currently improving (Schmidhuber 2008). On this account, the drive to compress better is not merely a useful modeling heuristic but the mechanism generating science, art, music, and humor alike, which makes it the clearest available target for the argument of this paper: if the relationship between coding deficit and distinguishability deficit is only proven in one direction and only at the extreme, then a theory that identifies the felt sense of understanding, or beauty, or interest with compression progress at every point along the curve, rather than only at the point where compression is complete, is claiming a great deal more than the formal relationship between the two quantities actually licenses. Sparse-coding and disentanglement objectives are frequently motivated by an appeal to compression as a proxy for the discovery of a domain's true generative factors, and post-hoc interpretability work often treats the discovery of a smaller, cleaner set of directions or features within a model's activations as evidence that those features constitute the model's

genuine ontology, on grounds close to Hillis's, and close to Schmidhuber's: the shorter description is taken to be the truer one, and the steeper the recent gain in shortness, the more interesting or better understood the underlying structure is taken to be. Two examples make the pattern concrete rather than merely characterized. The beta-VAE architecture increases a single compression pressure—a weighting term on the model's latent-code cost, pushed higher than in an ordinary variational autoencoder—specifically in order to produce latent representations described as more interpretable and as more faithfully capturing a domain's independent underlying factors of variation, with the increase in compression pressure offered as the mechanism by which the improvement in interpretability is obtained (Higgins et al. 2017). Sparse-autoencoder interpretability work on language models motivates the move to sparse, low-dimensional dictionaries of features in comparable terms: individual neurons are described as polysemantic and therefore poorly understood, while the sparse, more compressed feature directions recovered by the autoencoder are described as monosemantic and correspondingly more interpretable, with the compression step treated as what makes the improved interpretability available rather than as an independent, separately justified engineering choice (Bricken et al. 2023). In neither case is the inference from compression to understanding stated as an unproven assumption requiring a caveat; in both, it is the stated motivation for the method.

The apparatus of Sections 2 and 3 applies to this case without modification, and the case exposes a further difficulty the Forth case does not. It is worth being precise about what is being targeted before applying it. Many researchers working on sparse representations, mechanistic interpretability, and minimum-description-length methods already treat compression as a useful heuristic rather than as a theorem about understanding, and nothing in this paper is addressed to a practitioner who already holds the distinction Section 3 makes formal. The target is narrower and more common than any claim about what a field collectively believes: compression, or a reduction in a representation's apparent size, is routinely offered *as evidence for* deeper or more genuine understanding, in papers, talks, and informal argument alike, without the caveat that only one direction of the relevant relationship, and only its extreme case, has actually been established. It is the inference, not the community, that this paper is aimed at, and the inference survives even where individual practitioners would, if asked directly, decline to endorse it.

The Proxy Theorem's proven direction depends on a faithful-encoding assumption: that every distinction a representation expresses costs at least one unit of the representation's capacity to express. Distributed and superposed representations, of the kind large trained networks are now widely understood to use—in which a single direction in activation space participates in encoding more than one feature, and a single feature is encoded across more than one direction—are precisely representations for which faithful encoding, in the sense the theorem requires, need not hold. Where it fails, even the one direction of the Proxy Theorem that is actually proved no longer applies, and a reduction in a representation's measured or apparent description length carries no guaranteed relationship to its distinguishing capacity at all, in either direction. The inference from compression to understanding is, in this setting, not merely unproven past the point of perfect compression; it is not obviously licensed anywhere along the range where trained models actually operate.

None of this is an argument that compression is irrelevant to understanding, and the paper does not make that argument. A system that achieves zero coding deficit does thereby achieve zero distinguishability deficit, and that is a real, useful, and nontrivial guarantee at the extreme. The argument is narrower and, for that reason, more defensible: that the graded, continuous use of description length as a running proxy for understanding, throughout the actual working range of real representational systems rather than only at the unreachable extreme where both deficits vanish together, is not supported by the result

usually cited in its favor, and in the specific case of distributed neural representations may not be supported by any comparable result at all.

6 Curricula Must Widen

A further difficulty for any theory that identifies understanding, or interest, or progress with monotonic compression is that the environments in which learning actually occurs are not found in a state of nature; they are built, and built deliberately simple, for reasons that have nothing to do with the learner's understanding having reached some genuine informational floor. Each scaffold described in this section temporarily sacrifices environmental richness in order to preserve the distinctions a learner can currently acquire: not every distinction the environment could in principle present, but the subset a learner at the current stage can actually absorb, with the remainder deliberately withheld rather than lost. Children are not raised inside the full chaos of an unfiltered natural environment, in which every relevant variable changes at once and at every timescale simultaneously; they are raised inside environments deliberately constrained to present regularities slowly, one or a few at a time, precisely because a learner who cannot yet compress much of anything will fail to learn from an environment offering nothing but unstructured, simultaneous complexity. The same engineering shows up outside pedagogy proper, for reasons that are frankly economic rather than developmental. Standardized building materials—dimensional lumber cut to a small number of fixed sizes, sheet goods produced in uniform panels—constrain the built environment to a narrow vocabulary of regular shapes because mass production rewards uniformity, and the resulting simplicity of built spaces is a byproduct of manufacturing economics, not evidence that simple, rectilinear spaces are closer to the true shape of anything. Roads are built straight and their surfaces held deliberately unchanging across long distances in part because the vehicles traveling them have limited suspension travel and no capacity to adapt their gait to terrain the way a walking animal does; the road's low geometric entropy compensates for the vehicle's narrow tolerance, and says nothing about the terrain the road happens to cross.

In each case, the low entropy of the environment is a scaffold fitted to the current, limited distinguishing capacity of whatever is operating within it, not a report on the true complexity of the domain the environment sits inside. This matters for the compression-progress account directly, because a scaffold is only useful for as long as the learner has not yet outgrown it. Once a learner is no longer acquiring anything new from a fixed training environment, whether that learner is a child who has mastered a curriculum or a system whose compression of its current data has plateaued, further progress requires increasing the entropy of what is presented—widening $D_T(x)$ toward the distinctions the scaffold had been withholding—not continuing to compress what is already compressed as far as it usefully can be. A curriculum that only ever simplifies, and never widens, is a curriculum that terminates in a plateau rather than in understanding. This is difficult to reconcile with a theory that identifies the drive toward understanding with the drive toward compression progress at every point along the learner's trajectory, since the trajectory that theory describes is one of steadily increasing compression, while the trajectory that actually produces continued learning is one of alternating compression and deliberately reintroduced complexity.

A related and sharper difficulty appears within problem-solving itself, on timescales far shorter than a curriculum. Solving a problem often requires making it more complicated for a while before it can be made simple: introducing an auxiliary variable that has no simplifying role on its own but opens a path to one, embedding a problem in a higher-

dimensional space where a solution becomes visible before projecting the result back down, or otherwise moving, temporarily and deliberately, into what might be called a higher-dimensional logic relative to the problem's original statement. A compression-progress account, tracking description length as it evolves along the solver's trajectory, would register this intermediate expansion as a local reversal—a temporary loss of interestingness or beauty, on that account's own terms—precisely at the point where the solver is doing the genuine work that will shortly produce the solution. A theory of understanding that cannot distinguish a productive detour through greater complexity from a genuine loss of progress has mismeasured the trajectory it is trying to describe.

The curriculum-widening argument is not merely anecdotal; it has an independently developed empirical literature behind it that was arrived at for unrelated reasons. Developmental research on the explore–exploit trade-off argues that the distinctive length of human childhood is itself an evolved solution to a tension between exploring broadly and exploiting what is already known, with early cognition characterized by wide, high-entropy hypothesis search that is progressively narrowed only as the demands of goal-directed exploitation increase, and with young learners repeatedly found to explore more broadly than older learners and adults precisely where broad exploration is advantageous (Gopnik 2020). On this account childhood is not an inefficient, not-yet-compressed version of adult cognition working its way toward an adult's compression; it is a distinct cognitive strategy, favoring the preservation of distinctions a purely exploitative learner would discard, for exactly as long as the developmental trade-off favors exploration over efficiency. A learner whose only objective were continuous compression progress would converge on a narrow, exploited hypothesis space too early, at the first local efficiency gain available, and would never acquire the broader distinguishing capacity that later skilled performance depends on. Human development instead allocates an extended period in which exploration is favored over immediate efficiency, which is a second, independent line of evidence for the same structural claim Section 6's engineered examples make: that widening a curriculum, not narrowing it further, is what continued distinguishing capacity requires at exactly the point where a monotonic compression-progress account would expect further narrowing.

7 What Should Be Measured Instead

If coding length and distinguishing capacity can come apart as far as Sections 4 and 5 suggest, then a practice that measures the first while reporting conclusions about the second is measuring the wrong quantity whenever the two are not both pinned to zero. The correction this paper proposes is not a rejection of compression as a modeling virtue—a system with a smaller coding deficit is, all else equal, a more efficient one, and efficiency is its own legitimate goal—but a separation of that virtue from claims about understanding, which should be evaluated against the distinguishability deficit directly wherever the two quantities can be independently assessed: does the representation actually preserve the ability to separate the objects, cases, or outcomes that matter for the task, rather than merely occupying fewer bits while doing so.

This is a harder quantity to measure than description length, which is precisely why description length has been used as its substitute. The Forth case suggests one route past the difficulty: ask not how short a representation is, but what library of prior, admissible distinctions its shortness is reachable from, and whether that library, examined on its own terms, actually covers the distinctions the task requires. The machine-learning case suggests a second: ask whether a representation's apparent compactness survives an audit for superposition and faithful encoding before treating that compactness as evidence of

anything beyond itself. Neither route is as convenient as reading a bit count off a trained model or a line count off a Forth program. Convenience is exactly what the coding deficit offers, and exactly what the Deficit Proxy Theorem, read correctly rather than read as its popular simplification, does not license treating that convenience as. Appendices A through D restate the formal apparatus this recommendation rests on, so that the case for measuring distinguishing capacity directly can be checked against exactly what has, and has not, been proved about its relationship to coding length, rather than taken on the paper's word.

A Formal Definitions

For reference, the definitions used throughout the paper, restated together and without surrounding argument.

An *ontological triple* is $(X, \sim, T)$, where X is a set of objects, $\sim$ is an equivalence relation on X specifying observational indistinguishability, and T is an ontology: a description language or template hierarchy determining which descriptions of elements of X are actually reachable, as distinct from which descriptions merely exist in principle.

A description of $x \in X$ is *reachable within* T if it can be constructed using only the templates, operations, or primitives T makes available. The *reachable length* $L_T(x)$ is the length of the shortest description of x reachable within T. The *Kolmogorov-optimal length* $L^*(x) = K(x)$ is the length of the absolute shortest program capable of generating x, with no restriction to any particular ontology.

The *coding deficit* is $\delta_T(x) = L_T(x) - L^*(x)$, always non-negative since $L^*(x)$ is a minimum over all ontologies.

The *distinguishing capacity* $D_T(x)$ is the number of other objects separable from x by descriptions reachable in T. The *maximum distinguishing capacity* $D^*(x)$ is the largest such capacity attainable by any ontology. The *distinguishability deficit* is $\Delta_T(x) = D^*(x) - D_T(x)$, likewise always non-negative.

Faithful encoding is the assumption that every distinction a description language can express costs at least one bit of description length to express: no distinction is expressible for free.

B The Deficit Proxy Theorem

Theorem 1 (Deficit Proxy, restated)**.** *Under faithful encoding,* $\Delta_T(x) \leq C\,\delta_T(x)$ *for a system-dependent constant* $C > 0$*. In particular,* $\delta_T(x) = 0 \implies \Delta_T(x) = 0$*.*

Proof sketch. Under faithful encoding, each distinction inaccessible to T at x—each of the $\Delta_T(x)$ separations T fails to make that some ontology achieving $D^*(x)$ would make—would require at least one additional bit in T's description of x to flag, were T to be extended to express it. The number of such extensions needed is bounded by $\Delta_T(x)$, so the excess length $\delta_T(x)$ needed to close the gap is bounded below by $\Delta_T(x)$ up to the constant C scaling the maximum per-distinction cost, giving $\Delta_T(x) \leq C\,\delta_T(x)$.

For the special case: if $\delta_T(x) = 0$, then T already achieves the Kolmogorov-optimal description length for x. Under faithful encoding, every distinction that would cost a bit to express is therefore already expressible in T—an ontology paying the minimum possible length has no unspent length with which additional, unexpressed distinctions could be hiding—so no distinction is inaccessible: $\Delta_T(x) = 0$. □

[Full Deficit Correspondence, not established] Under faithful encoding, $\Delta_T(x) = 0 \iff \delta_T(x) = 0$.

The forward direction is Theorem 1 above. The reverse direction — that a distinguishability deficit of zero forces a coding deficit of zero — requires the further assumption that optimal coding uses every distinction available to it: that an ontology achieving maximum distinguishing capacity does so via a description that is also length-optimal, rather than via a longer description that happens to preserve every distinction without compressing them efficiently. No argument for this further assumption is offered in the source theorem, and this paper does not supply one either. It is exactly the assumption an inference from $\Delta_T(x) \approx 0$ to $\delta_T(x) \approx 0$ — the direction informal practice typically uses, since low cod-

ing deficit is what is actually observed and good distinguishing capacity is what is being inferred from it — would need, and does not have.

C Reachable Complexity, Formalized

The reachable length introduced in Appendix A is precisely the formal object informally called reachable complexity in Section 4:

$$L_T(x) = \min\{\, |d| : d \in T,\ d \mapsto x \,\},$$

the length of the shortest description d, drawn from the templates and operations available in T, that generates x. By definition of $L^*(x)$ as a minimum over all possible ontologies,

$$L_T(x) \geq L^*(x) = K(x),$$

with equality only when T happens to contain a description of x that is itself Kolmogorov-optimal — a reachability condition on T, not a general property of well-factored ontologies. A mature library can drive $L_T(x)$ very close to $L^*(x)$ for the objects it was built around, without this implying anything about $L_{T'}(x)$ for a differently-built library T', or about how close either library sits to $L^*(x)$ for objects outside the range its accumulated templates were shaped by. Forth's brevity, restated in this notation, is a claim about $L_T(x)$ for T equal to a specific, accumulated word library, not a claim about $L^*(x)$.

D Where the Two Deficits Diverge

The Deficit Proxy Theorem's asymmetry is not merely an absence of proof; the converse direction can fail outright. The following toy construction gives an ontology with a large coding deficit and a near-zero distinguishability deficit, showing the case Section 3 argues is left open is also, at least sometimes, actual.

[High coding deficit, low distinguishability deficit] Let $X = \{x_1, \ldots, x_n\}$ and let the true generating structure separate every pair: $D^*(x_i) = n - 1$ for each i. Let T describe each x_i by a fixed-width codeword of length $\lceil \log_2 n \rceil + m$ for some large padding constant $m > 0$ — inefficient relative to $L^*(x_i)$, which needs only $\lceil \log_2 n \rceil$ bits to distinguish n equally likely objects, so $\delta_T(x_i) \geq m$ for every i, growing without bound as m grows. But the codewords remain pairwise distinct for every value of m: padding a codeword does not merge it with any other, so T still separates every x_i from every x_j, giving $D_T(x_i) = n - 1 = D^*(x_i)$ and therefore $\Delta_T(x_i) = 0$ regardless of how large m, and hence $\delta_T(x_i)$, is made.

The construction is deliberately trivial — padding is not a realistic model of why real representations carry excess length — but its triviality is the point: nothing in the definitions of the two deficits rules out this combination, so no additional theorem is needed to show the combination is possible, only an example showing it actually occurs. A more realistic version of the same divergence is the ordinary condition of a verbose but complete encoding: any description language that never merges distinct objects, however wastefully it encodes them, has $\Delta_T(x) = 0$ by construction, whatever $\delta_T(x)$ happens to be. Section 4's Forth library and Section 5's superposed neural representations are offered as the substantive, non-toy versions of this same structural possibility; this appendix's role is only to confirm that the possibility is real rather than merely unproven-against.

References

[1] Flyxion. 2026. *Distinguishability Geometry*. Unpublished monograph.

[2] Flyxion. 2026b. *Compression After Expansion: On Skill, Misattribution, and the Geometry of Work*. Unpublished essay.

[3] Ziembik, Krzysztof A. 2025. "Forth: The Best Programming Language for the End of the World? A Case Study on A.D. 2044 (1991)." *Proceedings of the 36th ACM Conference on Hypertext and Social Media*.

[4] Rissanen, Jorma. 1978. "Modeling by Shortest Data Description." *Automatica* 14(5): 465–471.

[5] Schmidhuber, Jürgen. 2008. "Driven by Compression Progress: A Simple Principle Explains Essential Aspects of Subjective Beauty, Novelty, Surprise, Interestingness, Attention, Curiosity, Creativity, Art, Science, Music, Jokes." *arXiv preprint* arXiv:0812.4360. Short version published as "Simple Algorithmic Theory of Subjective Beauty, Novelty, Surprise, Interestingness, Attention, Curiosity, Creativity, Art, Science, Music, Jokes," *Journal of SICE* 48(1): 21–32, 2009.

[6] Higgins, Irina, Loïc Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. 2017. "beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework." *International Conference on Learning Representations (ICLR)*.

[7] Bricken, Trenton, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nicholas L. Turner, et al. 2023. "Towards Monosemanticity: Decomposing Language Models with Dictionary Learning." *Transformer Circuits Thread*.

[8] Gopnik, Alison. 2020. "Childhood as a Solution to Explore–Exploit Tensions." *Philosophical Transactions of the Royal Society B: Biological Sciences* 375(1803): 20190502.

[9] Anonymous. c. 2000. "The Forth's Dilemma, When Great Firms Cause New Technologies to Fail." Archived at the Internet Archive Wayback Machine, capture of http://www.msmisp.com/futuretest/Forth's_Dilemma.htm.