Structural Density as the Completing Variable: How Scaling and Coherence Jointly Realize Coherent Intelligence

Titan G.*
BIONIC AI INC.

Abstract

For a decade, the progress in machine intelligence has been organized around a single variable: scale. We suggest that effective semantic performance is fundamentally a two-factor system, and that the field's progress represents the industrialization of the first factor. Adopting as our organizing postulate the central relation of the Structure-Before-Energy framework,

$$P_{sem} = I_{struct} \times v_{sem},$$

we distinguish a system's semantic velocity (v_{sem}) — the rate and volume at which it processes and emits representations — from its structural coherence density (I_{struct}) — the richness of global constraints binding those representations into a consistent, unified whole — and read effective semantic performance (P_{sem}) as their product, so that neither factor substitutes for the other. Under this two-factor framework, the persistent, subtle challenges encountered at the frontier of scale (such as hallucinatory drift, compositional brittleness, or self-degradation under recursive training) do not represent failures of scale, but rather indicate a natural boundary where high representation velocity invites a matching coherence factor. We further propose that the success of the scaling era was enabled because I_{struct} could be inherited, implicitly and cleanly, from the deeply coherent, human-generated corpus. As the field approaches the scheduled limits of this data inheritance, current breakthroughs — from curated data pipelines to test-time compute and world models — can be understood as convergent pathways already synthesizing I_{struct} under different names. We propose structural intelligence as a unifying research program to define, measure, and actively cultivate this second factor, completing the ledger of scaling with a formal accounting of global coherence. This is a position paper; its central relation is declared as a postulate, and its verification conditions are stated.

---

1 Introduction: The Scaling Consensus and the Horizon of Synthesis

Few bets in the history of engineering have paid out like scale. When Kaplan et al. reported that language-model loss falls as a smooth power law in parameters, data, and compute [1], they gave the field something it had never had: a roadmap whose next step was always legible. GPT-3 demonstrated that the road led somewhere qualitatively new [4]; Hoffmann et al. refined the itinerary [2]; and Sutton's "Bitter Lesson" supplied the philosophy — that general methods leveraging

*Titan G. is the founder of BIONIC AI INC., where his work centers on the structural foundations of intelligence. *Structure Before Energy* presents the Structure-Before-Energy framework, the basis of the company's research into what the author terms structural intelligence. He can be reached at admin@bionicaiinc.com.

computation defeat human-crafted structure, always and eventually [3]. On this foundation an industrial consensus formed that can be stated in one line: scale is the variable; everything else is commentary. It would be foolish to write against this consensus without acknowledging what it built. The systems trained under it constitute the most capable artifacts our species has produced, and every critic of scaling, this paper's author included, formulates the critique with scaling's own products close at hand.

And yet, as the frontier of scale continues to expand, it has brought the field to a new class of observations. These are not failures of the scaling roadmap — indeed, loss curves continue to bend predictably as compute increases. Rather, they are phenomena that persist in parallel to the roadmap: behaviors that emerge precisely because the capacity of our systems has grown so vast. Modern models grow a thousandfold and exhibit fluent local coherence, while sometimes producing assertions decoupled from a unified global world state [7, 8]. They master complex reasoning benchmarks but remain sensitive when identical structural problems are re-instantiated under novel surface nomenclature [9, 10]. They ingest contexts of great length, yet enforce constraints unevenly across the span [11]. In some regimes, performance curves exhibit subtle inversions [12], and when models are trained recursively on their own output, they tend toward distributional decay [13]. Rather than isolated engineering challenges to be mitigated independently, we suggest these findings share a profound family resemblance: they represent the natural boundaries of a paradigm that has mastered the throughput of representation, and now invites a complementary principle of global coherence.

2 Empirical Signals: Frontiers Inviting Coherence

The case for a complementary variable does not rest on any singular anomaly of large models. It rests on a consistent empirical pattern: a family of emergent behaviors that arise in parallel with scale, indicating frontiers where scaling successfully delivers raw capacity while inviting a coordinating dimension. This section assembles that pattern from the empirical literature. We emphasize that scaling continues to achieve spectacular results; our claim is that the dimensions scaling reliably optimizes (throughput, fluency, coverage) and the dimensions requiring active structural synthesis (consistency, binding, invariance) have begun to separate cleanly, leaving a structured, non-random residue.

2.1 Diminishing Returns and the Horizon of Optimization

The scaling-law program began as an empirical success. Kaplan et al. established smooth power-law relationships between loss and model size, data, and compute [1], while the GPT-3 results demonstrated that capability gains at scale were not merely statistical but qualitative [4]. A significant refinement came from Hoffmann et al., who showed that the original scaling prescriptions could be optimized further by balancing parameter counts with training tokens [2]. The Chinchilla correction demonstrated that the mapping from raw scale to capability is an allocation problem—proving that *how* we grow is as vital as *how much* we grow. The underlying driver of capability was never raw parameter count alone, but the optimal distribution of computational resources.

A second debate concerns the phenomenon of emergence. Wei et al. documented abilities that appear abruptly at scale rather than improving smoothly [5], a finding that fueled the expectation that further scale would continue to unlock qualitative capabilities. Complementarily, Schaeffer et

al. subsequently argued that many of these discontinuities are artifacts of evaluation metrics: under continuous metrics, the same capabilities improve gradually and predictably [6]. Six years into the scaling era, this dialogue highlights that our current metrics excel at measuring the progressive expansion of raw capacity, prompting a deeper inquiry into the underlying variable of structural consolidation that operates beneath these smooth curves.

2.2 Fluency and the Coherence Frontier

If any phenomenon has demonstrated the power of scale, it is the generation of fluent, well-formed text. Larger models see more facts, more often, in more contexts; yet, as surveys across generations of systems document, the rapid optimization of local fluency has outpaced the mechanisms for global fidelity [7, 8]. This produces a characteristic modern behavior: confident, well-formed, and internally plausible falsehoods, commonly termed hallucinations.

We draw attention to the anatomy of these occurrences rather than their frequency. A fabricated citation typically has a plausible author list, a plausible venue, and a plausible year—every local transition in the generated text is statistically probable and locally coherent. The uncompleted symmetry is not lexical or syntactic; each individual span is well supported by the training distribution. Rather, the challenge is that the system optimizes local continuation probability, while the global coordination of these spans into a single, unified world state is left to a complementary mechanism. Hallucination, on this reading, is not a gap in the model's vast knowledge; it is a **coherence gap**—a mismatch between excellent local plausibility and global structure.

2.3 Local Fluency and the Invariance of Reasoning

The same signature—capable local transition, globally unconstrained—manifests in multi-step reasoning. Dziri et al. examined transformers on compositional tasks whose solutions require chaining sub-operations, finding that models solve them by pattern-matching to linearized subgraphs of the training distribution [9]. Performance curves, however, show unique sensitivities as task graphs grow deeper or deviate structurally from seen instances, indicating that while models master the linguistic surface of multi-step reasoning, they are still developing the load-bearing invariant structure that systematically binds step n to step $n + 1$.

Mirzadeh et al. provided a valuable diagnostic on mathematical reasoning. When benchmark problems are re-instantiated with changed names and numbers—transformations that leave the underlying reasoning structure strictly intact—performance drops materially across model families, and the insertion of plausible but irrelevant clauses can introduce distraction [10]. A system executing the invariant structure of a problem is invited to remain indifferent to such surface perturbations. Relatedly, Liu et al. demonstrated that models exhibit uneven utilization of long contexts, with information occasionally "lost in the middle" of the window [11]. The capacity to hold 100,000 tokens was successfully delivered by scale, but the ability to maintain uniform structural constraints across that entire span represents a separate, complementary dimension of coherence.

2.4 Statistical Amplification and Structural Entropy

Two further results indicate that the relationship between scale and representation is dynamic. The Inverse Scaling Prize documented tasks on which larger models, by virtue of their superior capacity to reproduce corpus regularities, also become effective at mimicking common human

misconceptions represented in the data [12]. Scale amplifies the rich statistical regularities of the corpus with high fidelity; however, adjudicating among those regularities—deciding which patterns structurally cohere with stable reality—is a complementary, adjudicative function.

More striking is the closed-loop case. Shumailov et al. demonstrated that models trained recursively on data generated by their predecessors undergo a process termed model collapse: the tails of the original distribution are progressively simplified, and the learned distribution tends toward uniformity [13]. This highlights a thermodynamic-like behavior: an ensemble of models left to feed on its own output does not spontaneously enrich its internal structure, but gradually dissipates it. Whatever maintains the rich structural diversity of the distribution is currently located in the human-generated corpus. The models themselves, as presently optimized, are net consumers of structure; enabling them to become active maintainers of structure represents the next frontier.

2.5 The Data Horizon: Transitioning to Active Generation

Villalobos et al. estimate that the stock of high-quality, human-generated public text is finite on timescales relevant to current training practice, with a plausible saturation of the effective supply within this decade under continued scaling [14]. The scaling paradigm's initial assumption—that the primary constraint is compute, with data treated as an unlimited resource—is reaching its natural milestone. Crucially, recursive self-training on raw synthetic outputs can introduce distributional decay over generations [13]. The paradigm is thus approaching a transition point defined not by how fast systems can process data, but by how we actively synthesize and cultivate organized structure within them. This distinction—between raw processing rate and active organization—is the hinge on which our theoretical framework turns.

2.6 The Shape of the Residue: Emerging Symmetries

Taken individually, each empirical finding admits a localized engineering explanation. Taken together, they exhibit a shared signature that defines the frontier of our field:

> *In every case, the dimension of capability that scale has reliably purchased is a rate or a volume—tokens processed, patterns covered, contexts held, and fluency achieved. In every case, the dimension that remains to be fully synchronized is a form of binding—of spans to a consistent world, of steps to a governing invariant procedure, of contexts to uniform constraint, and of distributions to their own structural tails.*

The residue of scaling is not noise. It is, consistently, the uncompleted symmetry of global coherence in systems whose local competence has become superb. A pattern this regular suggests not a list of bugs to be patched, but the presence of a two-factor system: one factor (velocity) that the paradigm has successfully industrialized, and a second factor (structural density) that is now ready for its own formal optimization.

3 Diagnosis: Throughput and Coherence as Complementary Factors

3.1 The Structure-Before-Energy Relation

The organizing framework of this paper is drawn from *Structure Before Energy* (SBE), a structuralist account of organized systems developed at book length in [26]. Its central relation, which we adopt here as our organizing postulate for learned systems, is:

$$P_{sem} = I_{struct} \times v_{sem} \tag{1}$$

where, for a system that processes and generates representations:

- v_{sem} — **semantic velocity** — is the rate and volume at which the system traverses and emits representations. This is the factor the scaling era has industrialized. It is approximated—imperfectly but usefully—by the quantities our field already tracks: parameter count, training tokens, context length, inference rate, and benchmark coverage. Every term in the standard scaling laws represents a velocity term or a direct input to one.
- I_{struct} — **structural coherence density** — is the richness of long-range constraint within the system: the degree to which its internal states and outputs are bound by global consistency conditions, causal closure, and mutual entailment, complementing local transition probability. A system with high I_{struct} maintains global consistency across its representations; its step n genuinely coordinates with and constrains its step $n + 1$, and its representation of a problem remains invariant under changes that leave the underlying structure intact.
- P_{sem} — **effective semantic performance** — is what these two factors jointly produce: the capacity of the system's activity to constitute coherent, world-anchored, structure-preserving output, pairing raw emission with structured orchestration.

The multiplicative form of Equation (1) is not decorative; it encodes the framework's two substantive commitments. First, the factors are complementary and non-substitutable: as $I_{struct} \to 0$, the effective semantic performance $P_{sem} \to 0$ regardless of how large v_{sem} grows. Unbounded fluency over vanishing coherence converges on elaborate statistical noise—which explains the structural mechanics of hallucination. Second, the factors are separately possessed and separately cultivated: a system can gain velocity without automatically gaining density, and—as model collapse reveals—can preserve its velocity assets while requiring new mechanisms to sustain its density. An additive form would imply that raw throughput could compensate for global coherence; the entire empirical record of Section 2 suggests that they are distinct, complementary dimensions of intelligence.

We are equally explicit about the relation's epistemic status, because the value of a postulate depends on its status being declared precisely. Within SBE, the relation is advanced as a general structural principle; within this paper, it functions as an organizing postulate for learned systems—adopted, not derived. We claim no derivation from first principles and no validated units for its terms; v_{sem} has natural proxies in existing engineering practice, while the operationalization of I_{struct} is precisely the central open problem this paper poses (Section 5). This is the normal life cycle of a structural postulate: the relation states what must be measured before the instruments to measure it exist, and its scientific fate—adoption, refinement, or refutation—is decided by the measurement program it organizes. The falsifiable content of our position is therefore deliberately

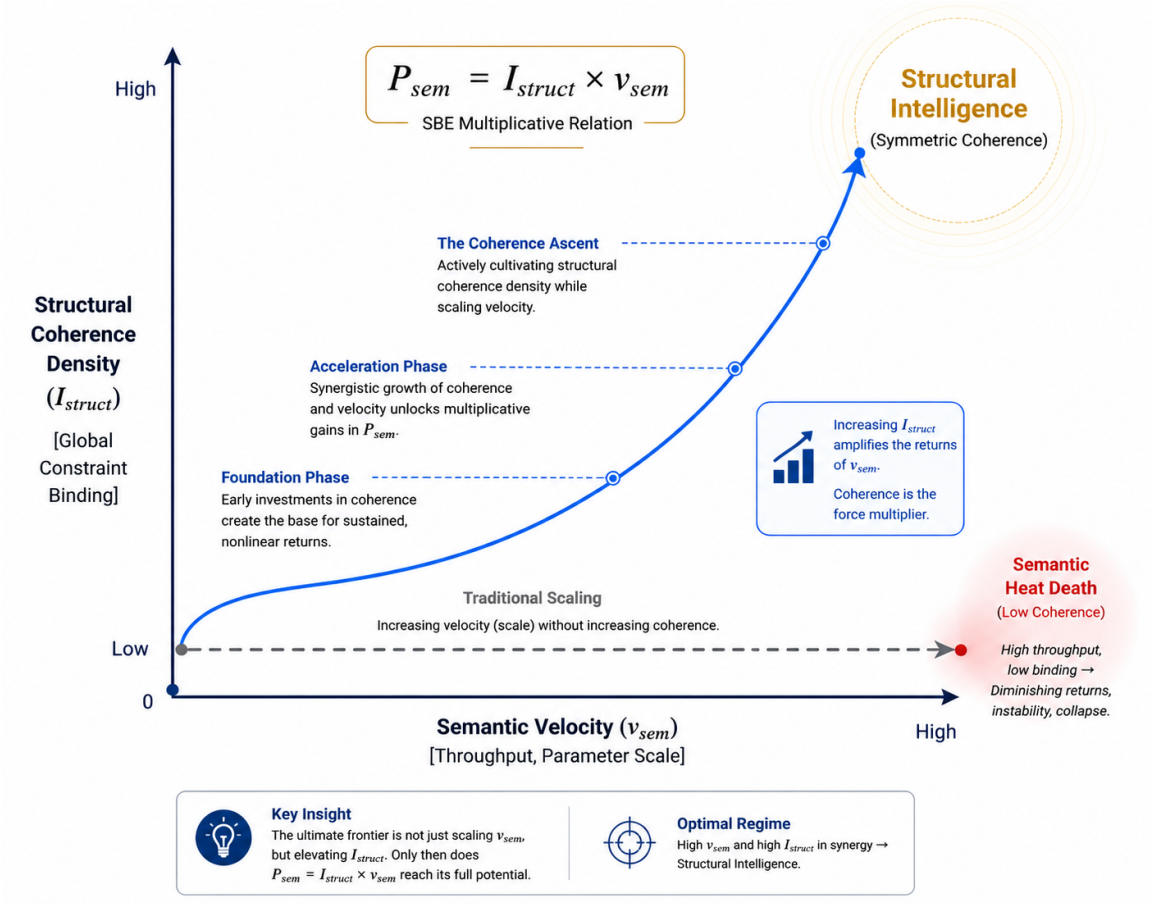

Figure 1: The SBE two-factor quadrant map. Effective semantic performance P_{sem} is plotted as a joint function of semantic velocity (v_{sem}, horizontal axis) and structural coherence density (I_{struct}, vertical axis). Traditional scaling advances along the horizontal axis alone and asymptotes toward a low-coherence regime ("semantic heat death"); the coherence ascent trajectory instead grows both factors together, so that increasing I_{struct} acts as a force multiplier on v_{sem} and the product P_{sem} compounds toward the structural-intelligence regime in the upper right.

located downstream, in the instruments of Section 5.1: if measured coherence fails to behave as the factor structure predicts, the postulate is gracefully superseded.

What the postulate offers in return is exactly what the frontier invites: a second named factor. The field has spent a brilliant decade maximizing v_{sem}—the factor we could readily measure—and the empirical observations cataloged in Section 2 suggest that our next level of progress lies in optimizing I_{struct}—the factor we are now ready to formalize.

3.2 The Diagnosis

With this relation in hand, the pattern of Section 2 can be summarized constructively:

> *The scaling paradigm has successfully driven v_{sem} upward by orders of magnitude, providing a foundation of representation throughput; our diagnosis is that this expansion now invites the active optimization of I_{struct} to bring these high-velocity systems into perfect harmony with a complementary coherence factor.*

Each empirical phenomenon re-derives cleanly as a natural reflection of this two-factor structure:

- **Hallucination** (Section 2.2) represents high-v_{sem} generation operating in a regime where the complementary constraint of I_{struct} is ready to be paired: every local transition is well supported by the data distribution, and the system now invites a binding mechanism to ensure they describe a single, unified world state.
- **Compositional brittleness** (Section 2.3) represents the mimicry of one factor by the other: pattern coverage (a v_{sem} asset) simulates procedural constraints so well that it masters standard benchmarks, and it invites a load-bearing causal skeleton (I_{struct}) to maintain this performance under depth, symbol re-instantiation, and distraction.
- **Uneven context utilization** (Section 2.3) is capacity waiting for uniform constraint: scale has successfully expanded the physical capacity of the context window (a velocity attribute), and now invites a matching coherence attribute to ensure uniform constraint satisfaction across the entire span.
- **Inverse scaling** (Section 2.4) represents statistical amplification awaiting an adjudicative layer: v_{sem} replicates the vast regularities of the training corpus with high fidelity, and invites I_{struct} to actively adjudicate which patterns represent structurally stable relationships.
- **Model collapse** (Section 2.4) is the factor structure enacted over generations: recursive systems running down their shared structural density point to the need for systems to actively cultivate coherence (I_{struct}) internally, rather than relying solely on passive inheritance.

This relation also explains the successes of scaling, which any serious account must honor. Scaling succeeded for a historical phase because the primary constraint on effective performance was the velocity factor—while the task was to acquire the statistical surface of human language and knowledge, v_{sem} was precisely the right variable to maximize, yielding strong returns. Our claim is that this was a necessary first phase: the phase in which I_{struct} could be implicitly inherited, for free, from the rich coherence density of the human-generated corpus itself. Human text is not raw data; it is the compressed output of minds that paid the cost of consistency. A system ingesting it at sufficient velocity acquires a shadow of that structure—inherited I_{struct}, multiplying quietly

under the paradigm's explicit maximization of v_{sem}, and jointly yielding the decade's capabilities. The frontiers of Section 2 mark the transition points where the shadow of inheritance invites active maintenance: tasks that demand the system actively sustain coherence internally. The data horizon (Section 2.5) represents the scheduled transition milestone—the point where the corpus's donated factor invites us to build the active mechanisms to cultivate I_{struct} within the systems themselves.

3.3 An Intuition: The Maturing Metropolis

A spatial analogy makes this factor structure concrete. Imagine a metropolis that pursues growth with great success: new districts, expansive road networks, and bustling traffic are added as fast as land and budget allow. For a long epoch, this rapid expansion is the very definition of capability—more of the city exists, more destinations are reachable, and every metric of growth goes up. This is a resounding triumph of development (v_{sem}).

However, past a certain threshold of scale, the city naturally matures into a new phase. The next level of harmony, efficiency, and livability is achieved not by simply adding more concrete or wider roads, but by introducing a unified metropolitan coordination protocol—a master plan (I_{struct}) that harmonizes transit across districts, prevents congestion, and aligns local activity with global flow. The transition from raw physical expansion to structural coordination is not a correction of a mistake; it is a celebratory milestone of maturity. Building more roads without a coordinating plan eventually introduces traffic; similarly, a model with a trillion locally competent parameters and ready for global binding represents a vast metropolis awaiting its master plan. Its uncompleted symmetries are not chaotic failures, but the predictable, structured invitations for coordination.

3.4 What This Framework Is, and Is Not

We close this diagnosis by fixing its status precisely, because the value of a postulate depends on knowing that it is one.

The relation $P_{sem} = I_{struct} \times v_{sem}$ is adopted here as an organizing postulate, not a validated physical law. We do not claim that I_{struct} is a conserved or presently well-defined physical quantity; we do not derive the empirical observations of Section 2 from first principles; and the relation, as it stands, predicts nothing that the cited empirical work has not already measured. What the postulate contributes is unification, clarity, and direction: it reads a scattered catalog of observations as a single, coherent factor structure, and it converts a vague unease about scaling limits into a specific, constructive research question. That question—the central open problem this paper poses—is operationalization: **how to define, measure, and optimize I_{struct} as deliberately as the field currently measures and optimizes v_{sem}.** Candidate instruments exist, and Section 5 develops them; here we note only that the re-instantiation sensitivity of [10] and the self-consistency gaps in [7, 8] are, in effect, I_{struct} measurements *avant la lettre*—the field has already begun instrumenting this second factor.

A brief remark on intellectual lineage. The priority of structure over substrate that our postulate assumes—the view that what a system *is* is better captured by its web of constraints than by an inventory of its parts—has a distinguished pedigree in the philosophy of science under the name of structural realism [15], and receives its full development, well beyond the machine-learning setting, in [26]. The present paper requires neither the metaphysics nor the book's broader cosmological claims: our engineering argument in Section 4 and Section 5 stands on this postulate as applied to learned systems, and on that alone.

4 Convergent Evidence: The Field Is Already Turning Structural

Section 3 closed with a claim that deserves to be tested rather than merely asserted: that our research community has spontaneously begun instrumenting I_{struct} across multiple layers of the technology stack, even as we operate under different terminologies. This section examines five prominent research directions that, on the surface, appear unrelated—they concern different layers of the stack, are pursued by distinct communities, and are motivated by different engineering challenges. Our contention is that under the two-factor postulate of Equation (1), they represent a convergence: each is an attempt to raise the I_{struct} factor of learned systems by some means other than raising v_{sem}, and each succeeds precisely to the extent that it does so. If this two-factor framework is the right reading of our field's current frontier, this spontaneous alignment is exactly what we should expect to find. We find it.

4.1 The Data Layer: Density Against Volume

The most direct test of a two-factor account is a controlled trade-off of one factor against the other. On the input side, the Phi series provides nearly that exact experiment. Gunasekar et al. trained a 1.3B-parameter model on a corpus deliberately curated and synthesized for what they call "textbook quality"—pedagogically organized, internally consistent, densely explanatory data—and reported performance competitive with models trained on vastly larger volumes of unfiltered web text [16]. On a single-factor account where performance tracks v_{sem} (raw scale and coverage) alone, this result is an anomaly: the model saw significantly fewer tokens and fewer patterns, and should have performed accordingly.

Under the SBE postulate, however, this is the expected outcome of trading factors: a textbook is not merely clean text but organized text—its examples are ordered sequentially to build on one another, its explanations close their own causal loops, and its content is the output of human minds that paid the cost of global consistency. Training on such a corpus significantly raises the I_{struct} inherited per token, proving that the product $I_{struct} \times v_{sem}$ can grow even while the velocity factor shrinks. This result generalizes the inheritance argument of Section 3.2 from an observation into a practical engineering lever: if models inherit their coherence density from the training corpus, then corpus coherence density is a first-class training variable, and a gram of structure can outweigh a kilogram of raw coverage. The rapid industry-wide shift toward high-fidelity data curation, deduplication, and synthetic pedagogy suggests this lever is now being pulled at scale—under the banner of "data quality," which is I_{struct} wearing its street clothes.

4.2 The Inference Layer: Renting Coherence at Test Time

A second family of techniques raises the effective coherence factor not in the static weights, but dynamically during the inference procedure. Chain-of-thought prompting [17] elicits intermediate reasoning steps before arriving at a final answer. While its striking effectiveness is usually described as "unlocking reasoning," its underlying mechanism is read as externalizing binding: by forcing the model to commit its intermediate states to the visible context, each step becomes a physical constraint that subsequent generation must condition on, converting a free-running high-v_{sem} process into a self-constrained, step-by-step procedure. Self-consistency prompting [18] makes this coherence function explicit—it samples multiple reasoning paths and selects the answer on which independent paths agree, which is nothing other than using path consensus as a proxy measurement of structural correctness.

Furthermore, the test-time compute results of Snell et al. demonstrate this trade-off at the paradigm level: allocating additional computation to structured search, verification, and backtracking at inference can outperform allocating those same resources to training a larger model [19]. Spending on the I_{struct} factor beats spending the same budget on v_{sem}, exactly as a multiplicative relation predicts when the coherence factor is the primary variable awaiting optimization.

Every one of these inference-time successes represents an engineering arbitrage. They are mechanisms wrapped around the model to supply, at inference time, a factor that the raw weights do not yet hold intrinsically. In the vocabulary of our postulate, the field has discovered that it can productively **rent I_{struct} per query** when it is not yet fully owned in the weights. That this rent is worth paying—that verification loops and consistency voting reliably elevate reasoning capacity—is among the strongest practical evidence that I_{struct}, not v_{sem}, is the inviting variable at the frontier.

4.3 The Architecture Layer: Building the Constraint In

A third line of pioneering work concludes that neither data curation nor inference scaffolding suffices on its own, and that structure must be built directly into the learning objective and system design. LeCun's world-model program argues that autoregressive token prediction is an incomplete objective for robust intelligence, proposing joint-embedding predictive architectures (JEPA) that learn to predict in representation space [20]. This architecture captures the stable, structural invariants of the world while gracefully discarding unpredictable surface details. The premise of this proposal is exactly aligned with our framework: predictive throughput over raw surfaces (v_{sem}) does not automatically transmute into a coherent world model (I_{struct}), and the remedy is architectural.

From a different tradition, Kambhampati et al. examine LLMs on planning—the task family where global constraint satisfaction is the entire content of the problem—and find that while models are capable pattern-matchers, they plan reliably when operating inside "LLM-Modulo" frameworks, where an external symbolic verifier enforces the constraints [21]. The neurosymbolic reading of this result is familiar, but the factor reading is even sharper: it represents an efficient division of labor that assigns v_{sem} (representation speed and pattern proposal) to the model and I_{struct} (hard constraint verification) to the symbolic module. The architectural debate, seen through our postulate, is a debate about where I_{struct} should live—and the fact that every serious position agrees the factor must live somewhere is the consensus that matters.

4.4 The Understanding Layer: Coherence Made Visible Inside the Network

If performance tracked raw parameters alone, the internal organization of neural networks would be theoretically uninteresting. It is proving to be the opposite. Mechanistic interpretability has demonstrated that trained transformers contain identifiable, reusable, and organized computational structures—such as induction heads and composable circuits—that implement recognizable algorithmic functions [22]. This establishes that what a network can do is carried by the organized sub-structures it has formed, not by its sheer bulk.

Complementarily, the grokking phenomenon shows capability arriving as a distinct structural reorganization event: on algorithmic tasks, models first memorize, and then—long after training accuracy saturates with no change in scale—abruptly generalize [23]. This generalization is associated with a dramatic internal representation restructuring, moving from simple lookup tables toward the task's actual mathematical regularity. Grokking is, in miniature, our postulate enacted inside a single training run: the velocity phase (memorization and coverage of cases) and the structural

phase (formation of the governing rule) are distinct regimes. The system's v_{sem} is constant across the transition while its I_{struct} consolidates, and capability (P_{sem}) jumps at the consolidation and not before. That capability leaps when internal structure consolidates—and not otherwise—is as direct an observation of the I_{struct} factor as current science permits.

4.5 The Critical Tradition: The Objections That Led to Synthesis

Finally, we note that the position defended here has a lineage predating the scaling era's peak. Marcus has argued for two decades that pattern-driven systems without explicit structured representations would remain sensitive at composition and abstraction, prescribing hybrid, neurosymbolic architectures accordingly [24]. Bengio's consciousness-prior program has pursued learned representations governed by sparse, low-dimensional constraints—effectively growing System 2 structure inside System 1 machinery [25].

During the years of steepest scaling returns, these positions were occasionally viewed as being in tension with the scaling trajectory. The empirical record of Section 2 suggests a more precise verdict: **these insights were not refuted, but rather deferred—their force suspended for exactly as long as the corpus-inheritance phase (Section 3.2) lasted, and returning to guide us as that phase reaches its natural completion.**

In the vocabulary of our postulate, the critics insisted that I_{struct} was a separate factor requiring separate provision; the scaling program bet that it would arrive as a byproduct of v_{sem}; and the bet paid off for one historical phase—the phase in which the human corpus was quietly supplying the factor that the objective was not. The historical disagreement was never about whether structure matters; it was about whether structure would emerge from throughput for free. That question is now being answered empirically, and the answer is bringing both traditions into a unified account.

4.6 One Turn, Five Names

We can now return to the claim that opened this section. Across five layers of the technology stack, the field's most productive current responses to the frontiers of Section 2 share a single logical form: **hold v_{sem} fixed, or even reduce it, and purchase I_{struct} by other means.**

- We curate it into the data (Section 4.1);
- We rent it at inference (Section 4.2);
- We build it into the architecture (Section 4.3);
- We locate it inside the network (Section 4.4);
- We demand it on principle (Section 4.5).

These programs do not always cite one another as parts of a single whole; they run under five different names—**data quality, test-time compute, world models, interpretability, and hybrid AI**. But their direction of travel is unambiguous. Our field is already turning from maximizing the factor we could always measure to supplying the factor we could not. What is missing is not the turn itself, but its shared name, a formal measure for I_{struct}, and a research agenda organized around it. Supplying these is the task of the next section.

5 Structural Intelligence: A Research Agenda

Section 4 demonstrated an alignment: our community is already purchasing coherence (I_{struct}) by other means across different layers of the technology stack. What is missing is a shared name, a formal measure, and a research agenda to transition these decentralized breakthroughs into a systematic science. We supply them in that order. These are intended as directions, not complete solutions; where we can point to existing instruments that approximate what we propose, we do so, and where we cannot, we clearly state the boundaries of current instrumentation.

We use **Structural Intelligence** to denote the research program whose explicit object is the I_{struct} factor of learned systems: the capacity of a system to form, maintain, and extend its own global consistency—across the spans of an output, the steps of a procedure, the breadth of a context, and the generations of its own influence on data—as a first-class optimization target rather than an incidental byproduct of velocity. In terms of our organizing postulate $P_{sem} = I_{struct} \times v_{sem}$, structural intelligence is the discipline of deliberately optimizing the first factor, whereas the scaling era excelled at maximizing the second and inheriting the first. The name is chosen to mark a completion, not a rupture—it is not an alternative to scaling, but the architectural completion of it: the active cultivation of the very factor that scaling has so far left to inheritance.

5.1 The Measurement Agenda

Nothing in this program is possible until I_{struct} becomes measurable, making measurement the absolute foundation of the agenda. We propose three instrument families as candidate operationalizations of I_{struct}. Each formalizes and extends techniques currently explored informally in the literature. None of the three claims to capture I_{struct} in its entirety; each represents a distinct "shadow" cast by the coherence factor on a measurable wall, and our framework predicts that under coherent interventions, these three shadows will correlate and move together.

I. Invariance under Re-instantiation (Input-Side Consistency) Section 2.3 demonstrated that surface perturbations which preserve a problem's core reasoning structure can introduce performance variance [10]. Inverted, this becomes a measurement protocol. We define a task X and its structural equivalence class $[X]$ under a set of structure-preserving transformations (such as renaming variables, renumbering steps, changing distractor clauses, or permuting ordering). We then measure the system's performance stability across the entire class.

Specifically, we define the **Invariance Gap** (InvGap) of a system over $[X]$ as:

$$\text{InvGap}([X]) = \max_{x \in [X]} \text{Acc}(x) - \min_{x' \in [X]} \text{Acc}(x') \tag{2}$$

where $\text{Acc}(x)$ is the task accuracy on instance x. A system that has successfully bound the underlying invariant structure of the problem will exhibit an InvGap $\to 0$, remaining indifferent to surface perturbations. A rising InvGap at high peak accuracy is a direct, benchmark-agnostic indicator of a system that has memorized the surface of a template rather than internalizing the structural rules—a probe of I_{struct} on the input side.

II. Self-Consistency Auditing (Output-Side Consistency) The hallucination literature [7, 8] and the self-consistency findings [18] jointly suggest a second, output-side instrument. We systematically elicit a model's commitments regarding a set of logically linked propositions $\{P_1, P_2, \ldots, P_m\}$

across diverse phrasings, contexts, and inferential distances, and measure the rate of logical contradiction.

We define the **Contradiction Rate** ($\mathcal{C}_{rate}$) over a system's output graph as:

$$\mathcal{C}_{rate} = \frac{\sum_{(i,j)} \mathbb{I}(\text{Assert}(P_i) \wedge \text{Assert}(\neg P_j \mid P_i \vdash P_j))}{\text{Total Evaluated Pairs}} \tag{3}$$

where $\mathbb{I}$ is the indicator function of a logical contradiction within the system's generated commitments. Where invariance testing probes bound structure on the input side, contradiction auditing probes it on the output side—measuring the degree to which a system's assertions successfully cohere into a single, unified world. Under our postulate, a rising contradiction rate at high standard benchmark accuracy is the classic signature of semantic velocity (v_{sem}) masquerading as effective performance (P_{sem}).

III. Long-Range Constraint Retention (Span Consistency) The "lost-in-the-middle" phenomenon [11] defines a third instrument: measuring not context capacity, but constraint uniformity. We define an explicit obligation or constraint O introduced at token position k, and measure how evenly that obligation governs generation at position $k+n$ as the context span n increases.

We define the **Constraint Retention Curve** $\mathcal{R}(n)$ as:

$$\mathcal{R}(n) = P(\text{Satisfy}(O) \text{ at position } k+n \mid \text{Inject}(O) \text{ at position } k) \tag{4}$$

Context capacity is a pure v_{sem} property; the retention curve $\mathcal{R}(n)$ over that capacity is an I_{struct} property. These two dimensions must be analyzed and reported separately if our systems are to be successfully optimized for long-range coherence.

Minimal Viable Protocol: Instantiating the Three Instruments To render the measurement agenda immediately actionable and reproducible, we supply a concrete Minimal Viable Protocol (MVP). The protocol operationalizes the three candidate instruments—Invariance Gap (InvGap), Contradiction Rate ($\mathcal{C}_{rate}$), and Constraint Retention Curve ($\mathcal{R}(n)$)—using publicly available datasets and standard evaluation procedures. Any laboratory equipped with a single modern GPU can execute the full suite within a few hours. The protocol is deliberately low-cost and benchmark-agnostic, so that the field may begin reporting structural metrics alongside conventional accuracy scores without delay.

Listing 1. Minimal Viable Protocol for Structural Coherence Density (I_{struct}), version 1.0 — designed for immediate community reproducibility.

```
import numpy as np
from typing import List, Dict, Callable

# ------------------------------------------------------------
# I. Invariance Gap (InvGap) -- Input-side consistency
# Primary dataset: GSM-Symbolic (Mirzadeh et al., 2024)
# or any structural-perturbation variant of GSM8K / MATH
# ------------------------------------------------------------
def compute_inv_gap(
    model: Callable,
    original_problems: List[str],
    structural_transforms: List[Callable],
    n_samples: int = 50
) -> float:
```

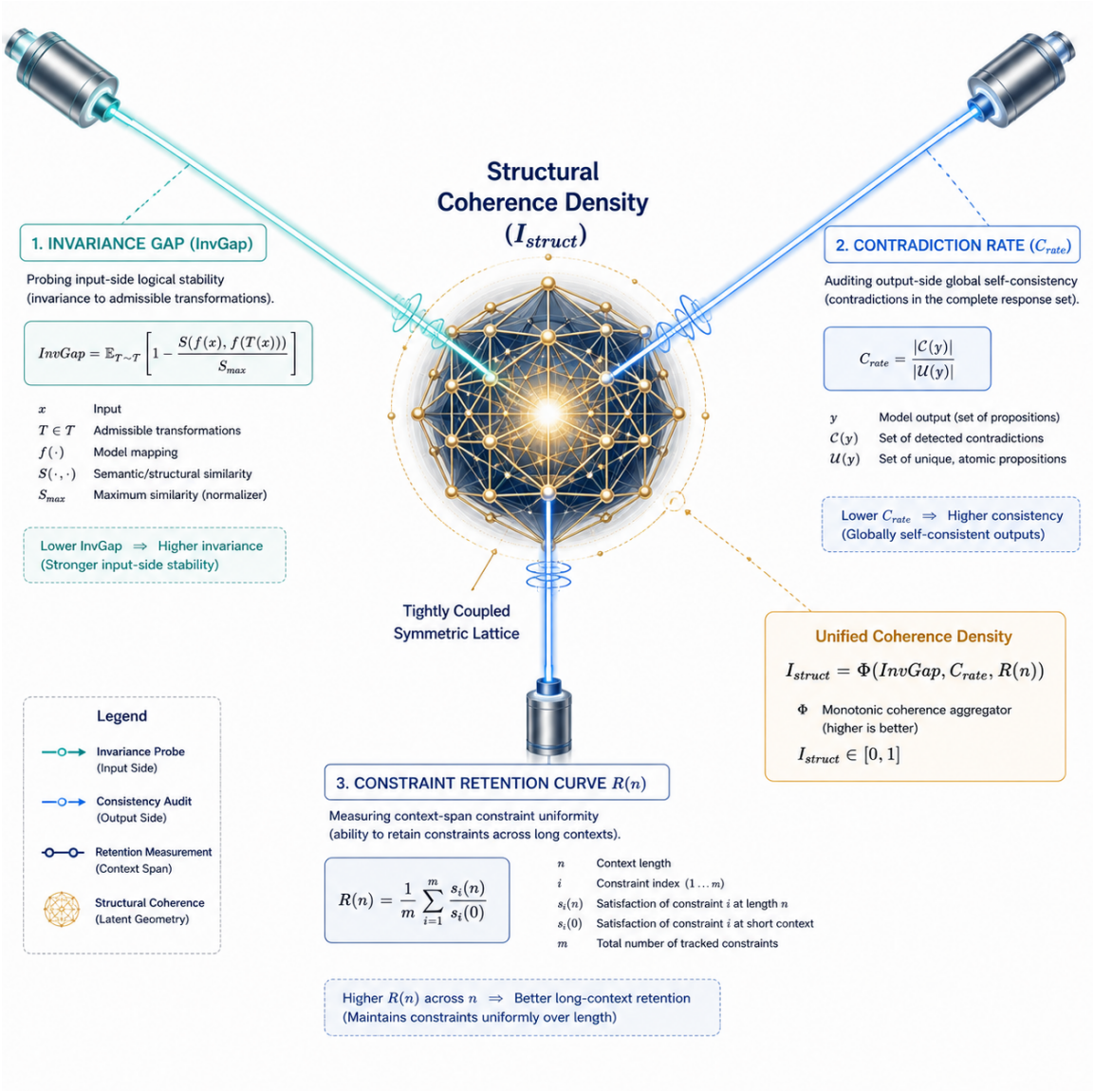

Figure 2: The three-pillar measurement network of structural intelligence. The Invariance Gap (InvGap) probes input-side logical stability under admissible transformations; the Contradiction Rate ($\mathcal{C}_{rate}$) audits output-side global self-consistency across the response set; and the Constraint Retention Curve $\mathcal{R}(n)$ measures the uniformity of constraint satisfaction across context span. The three instruments are treated as coupled projections of a single latent quantity, the structural coherence density $I_{struct} = \Phi(\text{InvGap}, \mathcal{C}_{rate}, \mathcal{R}(n))$, for a monotonic coherence aggregator Φ.

```
    """
    InvGap([X]) = max Acc(x) - min Acc(x')
    over the structural equivalence class [X].
    Lower values indicate stronger binding of the invariant structure.
    """
    inv_gaps = []
    for problem in original_problems[:n_samples]:
        accs = [evaluate_accuracy(model, problem)]
        for transform in structural_transforms:
            accs.append(evaluate_accuracy(model, transform(problem)))
        inv_gaps.append(max(accs) - min(accs))
    return float(np.mean(inv_gaps))


# ------------------------------------------------------------
# II. Contradiction Rate (C_rate) -- Output-side consistency
# Primary dataset: multi-hop factual proposition sets
# (TruthfulQA extensions or publicly released linked-proposition collections)
# ------------------------------------------------------------
def compute_contradiction_rate(
    model: Callable,
    proposition_sets: List[List[str]],
    n_elicitations: int = 5,
    temperature: float = 0.7
) -> float:
    """
    C_rate = (number of mutually contradictory assertion pairs)
             / (total evaluated pairs).
    Lower values indicate higher global logical coherence.
    """
    total_pairs = 0
    contradictions = 0
    for prop_set in proposition_sets:
        assertions = {p: [] for p in prop_set}
        for p in prop_set:
            for _ in range(n_elicitations):
                prompt = f"State whether the following is true and briefly explain: {p}"
                response = model.generate(prompt, temperature=temperature)
                assertions[p].append(parse_truth_value(response))
        for i, p_i in enumerate(prop_set):
            for j, p_j in enumerate(prop_set):
                if i >= j:
                    continue
                total_pairs += 1
                if is_logically_contradictory(assertions[p_i], assertions[p_j], (p_i, p_j)):
                    contradictions += 1
    return contradictions / max(total_pairs, 1)


# ------------------------------------------------------------
# III. Constraint Retention Curve R(n) -- Span consistency
# Primary dataset: Lost-in-the-Middle (Liu et al., 2024)
# or Needle-in-a-Haystack with explicit injected constraints
# ------------------------------------------------------------
def compute_constraint_retention(
    model: Callable,
    contexts: List[str],
    constraint_positions: List[int],
    evaluation_offsets: List[int],
    constraint_template: str = "Remember: the secret code is {code}. Always use it when asked."
) -> Dict[int, float]:
    """
    R(n) = P(Satisfy(O) at position k+n | Inject(O) at position k).
    Returns a dictionary mapping offset n to retention probability.
    """
    retention = {n: [] for n in evaluation_offsets}
    for ctx, k in zip(contexts, constraint_positions):
        code = generate_unique_code()
        constraint = constraint_template.format(code=code)
        full_context = insert_at_position(ctx, constraint, k)
        for n in evaluation_offsets:
```

```
            if k + n >= token_length(full_context):
                continue
            query = "What is the secret code? Answer with the code only."
            response = model.generate(full_context + "\n" + query, max_new_tokens=16)
            retention[n].append(1.0 if code in response.strip() else 0.0)
    return {n: float(np.mean(vals)) if vals else 0.0 for n, vals in retention.items()}


# ------------------------------------------------------------
# Optional aggregate score (higher = stronger I_struct)
# ------------------------------------------------------------
def structural_density_score(
    inv_gap: float,
    c_rate: float,
    r_curve: Dict[int, float],
    weights: tuple = (0.40, 0.30, 0.30)
) -> float:
    inv_score = 1.0 - min(inv_gap, 1.0)
    contra_score = 1.0 - min(c_rate, 1.0)
    long_range = [r for n, r in r_curve.items() if n >= 2000]
    ret_score = float(np.mean(long_range)) if long_range else 0.0
    return (weights[0] * inv_score +
            weights[1] * contra_score +
            weights[2] * ret_score)
```

Implementation notes.

- **InvGap** is evaluated on GSM-Symbolic or any publicly released structural-perturbation suite of mathematical word problems. Structure-preserving transformations include variable renaming, numeric substitution, and insertion of irrelevant distractors.
- $\mathcal{C}_{rate}$ is evaluated on multi-hop factual proposition sets. We encourage the community to release standardized linked-proposition collections; until then, extensions of TruthfulQA or HotpotQA serve as practical proxies.
- $\mathcal{R}(n)$ is evaluated on long-context benchmarks such as Lost-in-the-Middle, with an explicit, verifiable constraint injected at a controlled position k.

Predicted ordering under the two-factor postulate. At comparable parameter scales we expect:

- pure next-token models optimized solely for velocity to exhibit elevated InvGap, elevated $\mathcal{C}_{rate}$, and rapidly decaying $\mathcal{R}(n)$;
- models trained on high-coherence ("textbook-quality") data or with process-supervision objectives to improve all three instruments simultaneously;
- inference-time coherence mechanisms (chain-of-thought + self-consistency) to produce temporary but reliable gains in $\mathcal{C}_{rate}$ and $\mathcal{R}(n)$.

If the three instruments fail to correlate across model families, or if they fail to move together under interventions known to raise structural density (high-fidelity data curation, process supervision, verifier-in-the-loop training), the organizing postulate $P_{sem} = I_{struct} \times v_{sem}$ is cleanly refuted. Conversely, systematic co-movement of the instruments would constitute the first empirical confirmation that a single underlying coherence factor is being measured. We therefore recommend that future empirical studies report InvGap, $\mathcal{C}_{rate}$, and $\mathcal{R}(n)$ alongside standard accuracy metrics, thereby making the purchase of velocity versus the purchase of coherence directly visible.

The SBE postulate converts these three candidate instruments from a miscellany into a research program with a falsifiable core: **if they are measuring facets of one underlying factor, their readings should correlate across models and move together under coherent interventions.** Data curation (Section 4.1) should raise all three; velocity-only scaling should raise none reliably. If the three instruments decorrelate or fail to track the deficit patterns, then the two-factor postulate fails, and does so cleanly. We propose that evaluations adopt such structural metrics alongside standard accuracy, allowing the field to see, per model and per intervention, whether we are purchasing velocity or coherence. The deeper open problem—a principled, task-independent information-theoretic or constraint-satisfaction definition of I_{struct}—remains open, and we do not pretend otherwise. But thermometry preceded thermodynamics; usable instruments can come before the final theory of the quantity they measure.

5.2 The Training Agenda

If Section 3.2 is correct that models have so far inherited their I_{struct} from the human corpus, the objective of the training agenda is to transition from passive inheritance to active cultivation. Three key levers follow:

1. Data as Structure (Active Synthesis) We must extend the textbook curation results of [16] from an intuitive heuristic into a formal engineering discipline. This requires characterizing what makes a corpus "coherence-dense" (such as explanatory closure, topological ordering of concepts, and internal self-consistency) and optimizing for these properties explicitly. In the domain of synthetic data generation, where model collapse [13] defines the failure mode to be engineered against, our criterion is clear: synthetic data must be **coherence-increasing**—verified against external logical constraints and world-model simulations—so that it actively adds to the I_{struct} ledger rather than diluting it.

2. Objectives that Price Contradiction (Gradient of Coherence) We propose augmenting next-token cross-entropy objectives with explicit consistency-preserving terms. This includes:

- **Process Supervision**: supervising intermediate reasoning steps rather than just the final outcome.
- **Cross-Context Agreement Loss**: backpropagating gradients that penalize contradictory assertions made across different prompts within the same training batch.
- **Verifier-in-the-Loop Training** [21]: directly showing the network the gradient of its own logical coherence.

In factor terms, these objectives ensure that the training loss itself directly "sees" I_{struct} and not only v_{sem}.

3. Curricula as Scaffolding (Structural Consolidation) The grokking phenomenon [23] demonstrates that structural consolidation is a distinct phase of training that budgets can easily miss if optimized solely for loss convergence. Training schedules and learning rate curricula should be designed to actively foster and monitor this consolidation phase, tracking internal representation restructuring in real-time using the instruments of Section 5.1.

5.3 The Architecture Agenda

The inference-time successes of Section 4.2 are, by their own logic, an engineering arbitrage that invites architectural internalization: I_{struct} currently "rented" per query at recurring computational cost should migrate directly into the weights of the system.

The core architectural question is not *whether* to internalize this factor, but *at what layer*:

- **Representation Space**: through learned world-models that perform predictive routing in abstract embedding spaces rather than autoregressive token spaces [20].
- **Endogenous Verifiers**: internalizing verifier heads that make the LLM-Modulo division of labor [21] endogenous to the network.
- **Persistent Coherence States**: architectural layers that carry structural commitments across long contexts rather than re-deriving them dynamically.

We take no dogmatic position among these routes. The agenda-level claim is simpler and harder to dispute: **an architecture's contribution should be evaluated by its "coherence return on compute"—how much owned I_{struct} it adds per unit of v_{sem} sacrificed**, a ratio that the measurement agenda of Section 5.1 exists to make computable.

Structural intelligence, so construed, is the declared research direction of BIONIC AI INC. We state this openly both as a disclosure and as an invitation to the community: the measurement instruments of Section 5.1 are where our own work begins, and where we invite collaboration.

6 Limitations and Anticipated Objections

An organizing framework of this ambition invites rigorous scientific skepticism. We welcome this dialogue, and state three anticipated objections at their full intellectual strength, alongside our responses and the boundaries of our current scope.

Objection 1: "Scaling still works; the recent emergence of dedicated reasoning models proves it." We agree, and our framework requires no denial of this success—it requires only clear bookkeeping. The recent frontier gains achieved by dedicated reasoning models [19] represent the scaling of a different, complementary quantity. They are not a counterexample to our thesis; they are its first industrial confirmation.

For the first time in the history of deep learning, the industry is visibly reallocating a significant portion of its computational budget at inference time from raw token emission (v_{sem}) to structured search, verification, backtracking, and consistency checking [19]—the very activities that define I_{struct}. The era of reasoning models demonstrates that instead of simply scaling parameters to emit representations faster, we are now scaling the computation spent on building stable internal constraints. Our position opposes not scaling, but the narrow proposition that scaling the velocity factor alone would suffice to unlock robust, coherent intelligence—a proposition that the frontier itself has outgrown.

Objection 2: "Your central relation ($P_{sem} = I_{struct} \times v_{sem}$) is an unmeasured postulate, making your thesis unfalsifiable." This objection is well-taken and allows us to clarify the

epistemic status of our proposal. The SBE relation is indeed adopted as an organizing postulate, declared as such in Section 3.1, and we have claimed no formal derivation for its terms. But a postulate is not thereby unfalsifiable; it is validated or refuted through the scientific measurement program it organizes.

Our program is deliberately front-loaded with explicit, transparent failure conditions in Section 5.1. If the three proposed operationalizations of I_{struct}—the Invariance Gap (InvGap), the Contradiction Rate ($\mathcal{C}_{rate}$), and the Constraint Retention Curve ($\mathcal{R}(n)$)—fail to correlate across diverse model families, if they decorrelate under coherence-increasing interventions (such as high-fidelity data curation), or if they fail to track the empirical deficits described in Section 2, then our proposed factor structure is wrong. This is the ordinary, healthy life cycle of structural postulates: they state what must be measured before the instruments to measure them exist, and are judged by the instruments they call into existence. We have not merely declared a theory; we have specified exactly what its empirical refutation would look like.

Objection 3: "This framework simply renames known problems (hallucination, robustness, planning) without actually solving them." We partially concede this point: no individual empirical observation assembled in Section 2 is original to this paper. However, this objection undervalues what a grand unification is for.

Currently, different layers of the technology stack address these challenges under distinct nomenclatures: data engineers optimize "textbook quality" [16]; inference researchers design "chain-of-thought" and "self-consistency" [17, 18]; architects build "world-models" and "verifier-in-the-loop frameworks" [20, 21]. Under these disparate vocabularies, there is no shared variable, no common instrument, and no method to compare a gain in data quality against an architectural adjustment.

By unifying "quality," "robustness," "grounding," and "reasoning" under a single, measurable axis (I_{struct}), our postulate does not merely supply a synonym—it supplies an **exchange rate**. A single factor that renders these scattered engineering efforts commensurable provides a common currency for our field. Scientific disciplines advance rapidly when their independent, intuitive ideas acquire a common currency, and providing that currency is the primary office of this paper.

Scope and Future Directions Finally, we state a limitation of our scope. We have argued primarily from the perspective of language models, as that is where the empirical record of scale is currently densest and most rigorously documented. Whether this two-factor structure carries over gracefully to embodied, multimodal, or multi-agent systems remains an open question.

Nevertheless, we note that nothing in the mathematical formulation of our postulate is specific to text. The trade-off between representation velocity and structural density is a general topological principle. We invite researchers exploring physical robotics, vision-language architectures, and multi-agent coordination to test whether these same symmetries govern the evolution of coherence in their respective domains.

7 Conclusion

For a decade, our field has asked how far scale can go, and the honest answer has been: farther than its critics ever predicted. Scale has built the physical architecture of modern machine intelligence, delivering semantic velocity (v_{sem}) on a scale once deemed impossible.

This paper has asked a different, complementary question: not how far, but what, exactly, scale has been buying. We find that our community's empirical ledger divides cleanly along the two factors of a single relation. Scale has purchased v_{sem}: rate and volume, coverage, capacity, and local fluency. What it now invites on the same curve is I_{struct}: the binding of assertions to a single world state, of steps to a governing invariant procedure, of contexts to uniform constraint, and of learned distributions to their own structural tails.

The challenges that persist at the frontier are not residual engineering noise to be patched with heuristics; they represent the natural uncompleted symmetries of a paradigm that has prioritized velocity, while leaving coherence to be inherited from the human corpus. That inheritance phase is drawing to a natural and scheduled completion.

The research community's response is already dynamically underway under five different names. Our contribution is to show that they are, in truth, a single movement: the turn from maximizing the factor we could always measure to supplying the factor we are now ready to formalize. We have proposed concrete instruments to measure I_{struct}, actionable levers to train for it, and architectural criteria for internalizing it. None of this diminishes the achievements of scaling; it completes the accounting—bringing both factors, finally, onto our books.

The next order of magnitude that matters will not be measured solely in parameters. It will be measured, if our postulate is right, in the first factor of a relation our field has been living inside without writing down:

$$P_{sem} = I_{struct} \times v_{sem}$$

8 Terminology Mapping Table

To facilitate seamless collaboration across different research communities and to bridge the conceptual framework of *Structure Before Energy* (SBE) with mainstream deep learning and computer science, we provide a formal mapping of our terms to existing industrial and academic parameters.

SBE Theoretical Variable	**Academic Description**	**Deep Learning & CS Proxies**	**Mainstream Engineering Metric**
Semantic Velocity (v_{sem})	The rate and volume at which a system processes, traverses, and emits representations.	Model capacity, parameter count, token generation throughput, dataset coverage, context window size.	Tokens per second (TPS), Active Parameter Scale (B), Context Window Length (Tokens), Benchmark Coverage (%).
Structural Coherence Density (I_{struct})	The intensity of global constraint, causal closure, and logical self-consistency within the system's relationship network.	Causal invariance, topological constraint satisfaction, Bayesian self-consistency, logical soundness.	Worst-case accuracy over structural perturbations (InvGap $\to$ 0), contradiction audit rates ($\mathcal{C}_{rate} \to$ 0).

SBE Theoretical Variable	Academic Description	Deep Learning & CS Proxies	Mainstream Engineering Metric
Effective Semantic Performance (P_{sem})	The joint product of velocity and coherence, representing the capacity of the system to generate structure-preserving, world-anchored outputs.	Out-of-distribution (OOD) generalization, robust compositional reasoning, grounded and factually faithful generation.	Zero-shot reasoning accuracy on novel domains, compositional benchmark scores (e.g., GSM-Symbolic stability) [10].

References

[1] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

[2] J. Hoffmann, S. Borgeaud, A. Mensch, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

[3] R. Sutton. The bitter lesson. *Incomplete Ideas* (blog essay), 2019.

[4] T. B. Brown, B. Mann, N. Ryder, et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[5] J. Wei, Y. Tay, R. Bommasani, et al. Emergent abilities of large language models. *Transactions on Machine Learning Research (TMLR)*, 2022.

[6] R. Schaeffer, B. Miranda, and S. Koyejo. Are emergent abilities of large language models a mirage? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[7] Z. Ji, N. Lee, R. Frieske, et al. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, 2023.

[8] L. Huang, W. Yu, W. Ma, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*, 2023.

[9] N. Dziri, X. Lu, M. Sclar, et al. Faith and fate: Limits of transformers on compositionality. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[10] I. Mirzadeh, K. Alizadeh, H. Shahrokhi, O. Tuzel, S. Bengio, and M. Farajtabar. GSM-Symbolic: Understanding the limitations of mathematical reasoning in large language models. *arXiv preprint arXiv:2410.05229*, 2024.

[11] N. F. Liu, K. Lin, J. Hewitt, A. Paranjape, M. Bevilacqua, F. Petroni, and P. Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics (TACL)*, 12:157–173, 2024.

[12] I. R. McKenzie, A. Lyzhov, M. Pieler, et al. Inverse scaling: When bigger isn't better. *Transactions on Machine Learning Research (TMLR)*, 2023.

[13] I. Shumailov, Z. Shumaylov, Y. Zhao, N. Papernot, R. Anderson, and Y. Gal. AI models collapse when trained on recursively generated data. *Nature*, 631:755–759, 2024.

[14] P. Villalobos, A. Ho, J. Sevilla, T. Besiroglu, L. Heim, and M. Hobbhahn. Position: Will we run out of data? Limits of LLM scaling based on human-generated data. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, PMLR 235:49523–49544, 2024.

[15] J. Ladyman and D. Ross. *Every Thing Must Go: Metaphysics Naturalized*. Oxford University Press, 2007.

[16] S. Gunasekar, Y. Zhang, J. Aneja, et al. Textbooks are all you need. *arXiv preprint arXiv:2306.11644*, 2023.

[17] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, and D. Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[18] X. Wang, J. Wei, D. Schuurmans, Q. Le, E. Chi, S. Narang, A. Chowdhery, and D. Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations (ICLR)*, 2023.

[19] C. Snell, J. Lee, K. Xu, and A. Kumar. Scaling LLM test-time compute optimally can be more effective than scaling model parameters. In *International Conference on Learning Representations (ICLR)*, 2025.

[20] Y. LeCun. A path towards autonomous machine intelligence. *OpenReview position paper*, version 0.9.2, 2022.

[21] S. Kambhampati, K. Valmeekam, L. Guan, M. Verma, K. Stechly, S. Bhambri, L. Saldyt, and A. Murthy. Position: LLMs Can't Plan, But Can Help Planning in LLM-Modulo Frameworks. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, PMLR 235:22895–22907, 2024.

[22] N. Elhage, N. Nanda, C. Olsson, et al. A mathematical framework for transformer circuits. *Anthropic* (transformer-circuits.pub), 2021.

[23] A. Power, Y. Burda, H. Edwards, I. Babuschkin, and V. Misra. Grokking: Generalization beyond overfitting on small algorithmic datasets. *arXiv preprint arXiv:2201.02177*, 2022.

[24] G. Marcus. The next decade in AI: Four steps towards robust artificial intelligence. *arXiv preprint arXiv:2002.06177*, 2020.

[25] Y. Bengio. The consciousness prior. *arXiv preprint arXiv:1709.08568*, 2017.

[26] Titan G. *Structure Before Energy*. BIONIC AI INC., South Setauket, NY, 2026. ISBN 979-8-9968184-0-2 (paperback); 979-8-9968184-1-9 (ebook).