Language Models as Reflexively Autocatalytic Constraint Systems

Armando Vieira

Abstract

Large language models (LLMs) are trained by minimising an external prediction error, usually next-token cross-entropy. This description is operationally complete at the level of optimisation but incomplete as an account of the internal organisation that supports reasoning, in-context learning, recovery from error, and robust transfer. It also does not fully explain three recurrent observations: capabilities that appear in clusters, highly anisotropic brittleness under small perturbations, and substantial differences in capability between models with similar validation loss. We develop a formal interpretation of LLMs as reflexively autocatalytic and food-generated (RAF-like) systems embedded in a dynamic generative constraint field. In Generative Closure Theory (GCT), a capability becomes robust when task-relevant constraints become mutually supporting, temporally persistent, selectively sensitive to relevant evidence, and recoverable after disruption. We define behavioural and mechanistic closure measures, propose a training objective that acts directly on internal constraint states, and specify a reproducible experimental programme based on checkpoint scaling, matched-loss comparisons, perturbation recovery, isomorphic transfer, and causal activation interventions. The central empirical claim is falsifiable: closure-sensitive measures should predict robust capability onset beyond model size, compute, validation loss, and unperturbed task accuracy.

1 Introduction

Autoregressive transformers are normally defined through the conditional distribution

$$p_\theta(x_{t+1} \mid x_{\le t}), \tag{1}$$

and trained by minimising

$$\mathcal{L}_{\mathrm{LM}} = -\sum_t \log p_\theta(x_{t+1} \mid x_{\le t}). \tag{2}$$

This framework has produced reliable scaling laws relating cross-entropy loss to model size, data, and compute [3, 4]. It correctly describes the model's external training signal and the probability distribution exposed at its output. However, an external loss is a scalar summary over many contexts and tokens. It does not uniquely determine which internal variables are represented, how long they persist, whether they interact causally, or whether a correct trajectory can be reconstructed after disruption.

This limitation matters because LLM capabilities do not behave like a single smooth scalar. New abilities may become visible over a narrow range of scale [5]; some apparent discontinuities can be produced by nonlinear evaluation metrics [6]; and delayed generalisation can occur long after training accuracy has saturated, as in grokking [7]. Mechanistic work further shows that sharp behavioural changes may coincide with the formation of identifiable internal circuits. Induction heads, for example, appear near a sudden increase in in-context learning in controlled transformer models [8]. Causal tracing, activation patching, and sparse feature decomposition also indicate that output behaviour can depend on internal computational structure that is not visible from validation loss alone [9, 10].

Generative Closure Theory (GCT) proposes an intermediate explanatory level between parameters and behaviour. The model learns a high-dimensional field of constraints: entity identities,

relational rules, causal expectations, temporal commitments, task goals, stylistic requirements, and intermediate results. During generation, subsets of these constraints can become mutually supporting and self-maintaining. We call this state *generative closure*. The concept is inspired by reflexively autocatalytic and food-generated (RAF) systems, originally developed to formalise collectively self-sustaining reaction networks [13, 14, 15]. The mapping is organisational rather than chemical: internal products of computation facilitate subsequent transformations, while the prompt, context, memory, and external tools provide a task-dependent food set.

The main contribution is not a rejection of probabilistic modelling. Probabilities remain the model's output representation and cross-entropy remains an effective training signal. GCT instead asks what internal organisation makes those probabilities robustly meaningful. It predicts that scale changes capability by increasing not only representational capacity, but also the density, persistence, and recoverability of interacting constraints.

2 Why External Loss Is Not Enough

Equation (2) evaluates whether the next token receives high probability under teacher forcing. It is indifferent to many differences in internal implementation. Two models can assign similar probabilities to a benchmark corpus while relying on different mechanisms: one may maintain a stable entity-relation state, while another may reconstruct answers from local lexical cues. These strategies can be nearly indistinguishable on in-distribution loss yet diverge sharply under renaming, distractors, longer dependency chains, or contradictory context.

This creates a distinction between *external adequacy* and *internal organisation*. External adequacy is measured by loss, preference scores, or final-answer accuracy. Internal organisation concerns whether task-relevant constraints are:

(i) represented in a form that generalises across surface changes;

(ii) maintained across time and irrelevant intervening tokens;

(iii) integrated with other constraints rather than stored as isolated associations;

(iv) selectively updated when a premise changes;

(v) recoverable after noise, contradiction, or an incorrect intermediate state;

(vi) grounded by evidence external to the model's self-consistent linguistic trajectory.

The distinction is analogous to the difference between fitting an energy function and characterising the collective phases supported by its interactions. A scalar objective can drive the formation of structure without itself describing that structure. GCT therefore treats validation loss as necessary but not sufficient for explaining capability.

3 The RAF Interpretation of LLM Computation

A chemical reaction system is commonly written as

$$\mathcal{Q} = (X, \mathcal{R}, C, F), \tag{3}$$

where X is a set of molecule types, $\mathcal{R}$ a set of reactions, $C \subseteq X \times \mathcal{R}$ a catalysis relation, and $F \subseteq X$ an externally supplied food set. A subset $\mathcal{R}' \subseteq \mathcal{R}$ is RAF when it is both food-generated and reflexively autocatalytic [13].

For an LLM episode, we propose the following correspondence:

$$X \longleftrightarrow \text{latent features, active constraints, and intermediate states,} \tag{4}$$

$$\mathcal{R} \longleftrightarrow \text{attention, MLP, routing, retrieval, and tool-mediated transformations,} \tag{5}$$

$$C \longleftrightarrow \text{causal facilitation between representations and transformations,} \tag{6}$$

$$F_t \longleftrightarrow \text{prompt, context, retrieved evidence, memory, and tool outputs.} \tag{7}$$

Let $\mathcal{X}_t = \{c_1, \ldots, c_n\}$ be the active task-relevant constraints at step t, and let $\mathcal{R}_t = \{r_1, \ldots, r_m\}$ be the transformations available in the current computation. A transformation

$$r_j : S_j \to P_j \tag{8}$$

maps active source representations S_j into products P_j. The catalytic relation

$$\mathcal{K}_t \subseteq \mathcal{X}_t \times \mathcal{R}_t \tag{9}$$

records whether a representation causally facilitates a transformation. Facilitation can be estimated by activation patching, causal tracing, attention-path interventions, or changes in transformation success under feature ablation.

A task-dependent subset $\mathcal{R}'_t$ is RAF-like when:

1. **Food generation:** all representations required by $\mathcal{R}'_t$ can be reached from F_t through transformations within $\mathcal{R}'_t$;
2. **Reflexive support:** every transformation in $\mathcal{R}'_t$ is facilitated by at least one representation generated within the same organisation.

A reasoning chain is therefore not modelled merely as a sequence of probable tokens. Premises enable relational operations; intermediate conclusions constrain later operations; inconsistency signals activate repair; and the repaired state supports continued progress. The system temporarily regenerates the conditions of its own coherent continuation.

4 Generative Closure Formalism

At step t, define the active constraint graph

$$G_t = (V_t, E_t, W_t), \tag{10}$$

where V_t contains active constraints, E_t directed influence relations, and $W_t = (w_{ij})$ their causal strengths. Let $a_i(t)$ be the activation of constraint c_i. A soft internal-support measure is

$$\Gamma_t = \frac{1}{|V_t|} \sum_{i \in V_t} \sigma \left(\sum_{j \in V_t} w_{ji} a_j(t) - \tau \right), \tag{11}$$

where σ is a logistic function and τ a support threshold. Γ_t is large when active constraints are sustained by other active constraints rather than being independent responses to the most recent token.

Internal support alone is not sufficient. A hallucination may be linguistically stable and internally consistent while being externally false. GCT therefore separates five dimensions.

4.1 Constraint persistence

Let $q_v(d)$ denote the accuracy with which variable v can be decoded or behaviourally recovered d tokens after its last explicit mention. Define

$$d_{1/2}(v) = \min \left\{ d : q_v(d) \leq q_{\text{chance}} + \frac{q_v(0) - q_{\text{chance}}}{2} \right\}. \tag{12}$$

A long half-life indicates that a constraint remains functionally available rather than merely detectable at the moment it is introduced.

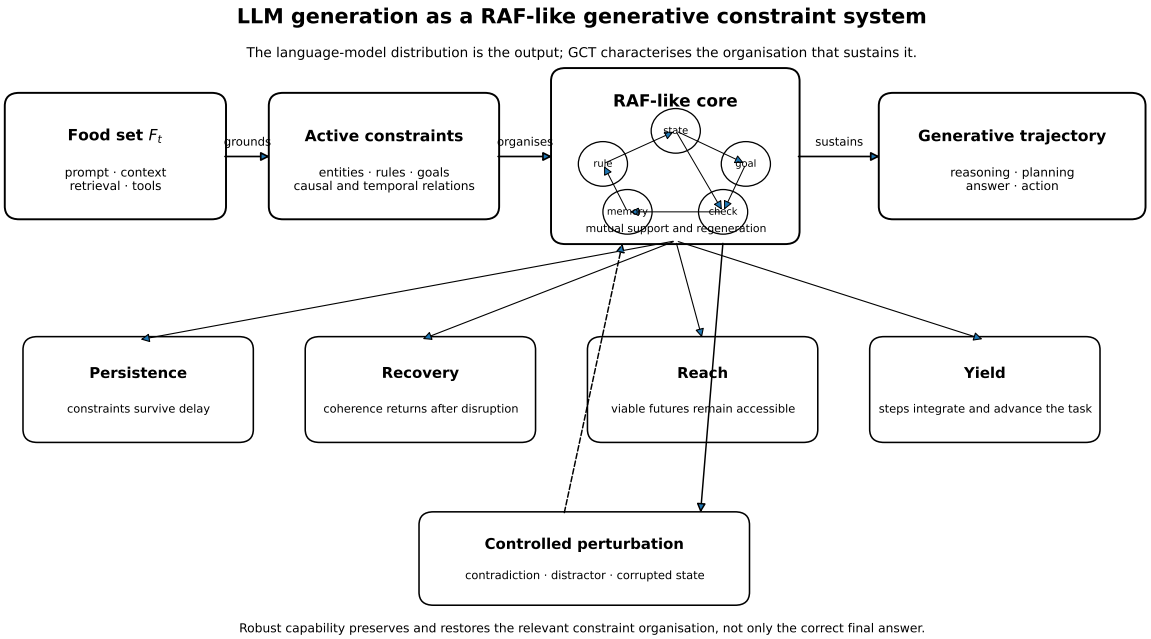

Figure 1: Conceptual mapping of LLM generation onto a RAF-like generative constraint system. A task-dependent food set activates constraints whose mutual interactions sustain a generative trajectory. GCT evaluates this organisation through persistence, recovery, reach, and yield.

4.2 Perturbation recovery

Let $\widetilde{x}$ be a context in which a contradiction, distractor, or corrupted intermediate state is inserted at step t. Recovery should be defined relative to a valid constraint-equivalence class, not to an identical token trajectory. Let $s_u^\star$ be the correct latent task state at later step u, and let $\widehat{s}_u(\widetilde{x})$ be the decoded state. Define

$$T_{\mathrm{rec}} = \min\left\{k : d\big(\widehat{s}_{t+k}(\widetilde{x}), s_{t+k}^\star\big) < \epsilon\right\}, \tag{13}$$

and

$$\rho = \Pr(\text{valid recovery}) \exp(-T_{\mathrm{rec}}/\kappa)(1 - E_{\mathrm{res}}), \tag{14}$$

where E_{res} is residual contamination by the perturbation.

4.3 Basin geometry

Let $\mathcal{P}_\tau$ be a family of perturbations with magnitude τ, including renaming, paraphrasing, distractor insertion, premise reordering, and bounded counterfactual changes. The behavioural basin volume is

$$B(\tau) = \Pr_{p \sim \mathcal{P}_\tau}[\text{task constraints remain satisfied}], \tag{15}$$

with basin width

$$w_B = \sup\{\tau : B(\tau) \geq \eta\}, \tag{16}$$

for a selected success threshold η. This makes brittleness a geometric property rather than a collection of isolated failures.

4.4 Reach and yield

Let $\Pi_H(t)$ be the set of viable future trajectories over horizon H that satisfy active constraints. Define

$$R_t = \log\left[\sum_{\pi \in \Pi_H(t)} \exp\{-\beta\, \mathrm{cost}(\pi)\}\right]. \tag{17}$$

Reach measures structured future possibility. Let Φ_t denote task progress, reduction of relevant inconsistency, or increase in verified constraint satisfaction. Yield is

$$Y_t = \Phi_{t+1} - \Phi_t. \tag{18}$$

High reach without yield corresponds to undirected openness; high yield with vanishing reach corresponds to premature closure.

4.5 External grounding

Let C_{int} measure internal consistency and V_{ext} external validity. The false-closure or grounding gap is

$$D_{\text{ground}} = C_{\text{int}} - V_{\text{ext}}. \tag{19}$$

A confident hallucination is predicted to have high internal closure but a large grounding gap.

5 Emergent Properties Under GCT

5.1 Clustered emergence

Capabilities such as entity tracking, compositional rule induction, in-context learning, contradiction detection, and tool use depend on overlapping subconstraints. If these components are initially fragmented, scale produces gradual local improvements without stable capability. When their interactions become sufficiently dense and mutually supportive, several tasks can become accessible together. GCT therefore predicts that capability-onset distance should be smaller for tasks sharing more subconstraints, after controlling for ordinary task difficulty.

This mechanism is more specific than saying that a larger probability model performs better. It identifies a shared organisational substrate. The appearance of induction heads near a sharp increase in in-context learning provides a precedent for linking behavioural onset to circuit formation [8], although GCT generalises the idea from one circuit type to prompt-dependent networks of constraints.

5.2 Brittleness geometry

An acquired capability may initially occupy a narrow basin. Familiar wording activates the correct organisation, while symbol renaming, irrelevant text, premise reversal, or a misleading local cue pushes the model into a different trajectory. Further training can enlarge the basin even when central-task accuracy changes little. GCT therefore predicts that robustness measures such as $B(\tau)$ and ρ may scale differently from ordinary accuracy.

This also reconciles genuine organisational change with the metric-artifact critique of emergence [6]. A benchmark score may jump because of thresholding even when the underlying response surface changes smoothly. GCT does not require every observed jump to be a phase transition. It predicts that the evolving basin geometry and recovery dynamics reveal whether a substantive organisational change lies beneath the score.

5.3 Loss–capability dissociation

Validation loss averages local predictive performance over a broad distribution. A comparatively rare multi-step capability may contribute negligibly to total loss. Two checkpoints can therefore have similar cross-entropy while differing in constraint half-life, isomorphic transfer, or contradiction recovery. GCT predicts that adding closure metrics to a capability model should explain variance not captured by loss:

$$\text{Capability} \sim \text{Loss} + \log N + \text{Compute} + \rho + w_B + d_{1/2}. \tag{20}$$

5.4 Delayed generalisation

Grokking shows that test generalisation can improve abruptly after training performance has saturated [7]. Under GCT, prolonged optimisation reorganises previously memorised associations into a more compact and mutually constraining structure. The key empirical prediction is not merely delayed accuracy, but a preceding increase in invariance, persistence, and recovery. Closure measurements should therefore provide progress indicators before final generalisation becomes visible.

6 A GCT-Oriented Training Objective

The proposed objective acts on internal task states as well as output tokens:

$$\begin{aligned}\mathcal{L}_{\mathrm{GCT}} = \mathcal{L}_{\mathrm{LM}} + \lambda_{\mathrm{inv}}\mathcal{L}_{\mathrm{inv}} + \lambda_{\mathrm{state}}\mathcal{L}_{\mathrm{state}} + \lambda_{\mathrm{rec}}\mathcal{L}_{\mathrm{rec}} \\ + \lambda_{\mathrm{dep}}\mathcal{L}_{\mathrm{dep}} + \lambda_{\mathrm{ground}}\mathcal{L}_{\mathrm{ground}}.\end{aligned} \tag{21}$$

Let $H_\theta(x)$ denote hidden activations at selected layers and positions, and let q_ϕ be an auxiliary decoder mapping activations to a structured task state $s(x)$.

State loss.

$$\mathcal{L}_{\mathrm{state}} = \mathbb{E}_x\, \ell(q_\phi(H_\theta(x)), s(x))\,. \tag{22}$$

This makes entity relations, active rules, temporal scope, and unresolved contradictions explicitly recoverable during training.

Structural invariance. Let g be a structure-preserving transformation, such as arbitrary renaming, and let P_g be the corresponding permutation of the structured state. Then

$$\mathcal{L}_{\mathrm{inv}} = \mathbb{E}_{x,g}\, \|q_\phi(H_\theta(gx)) - P_g q_\phi(H_\theta(x))\|_2^2\,. \tag{23}$$

The model is trained to preserve relational structure while allowing surface form to change.

Recovery loss. For a perturbed context $\widetilde{x}$,

$$\mathcal{L}_{\mathrm{rec}} = \mathbb{E}_{\widetilde{x}} \sum_{u=t+1}^{t+H} \omega_u\, \ell(q_\phi(H_{\theta,u}(\widetilde{x})), s_u^\star)\,. \tag{24}$$

The target is a valid state trajectory, not exact reproduction of the unperturbed wording.

Dependency-selective update. Let δ alter one premise and M_δ identify state variables that should change. Then

$$\begin{aligned}\mathcal{L}_{\mathrm{dep}} = \ell\Big(q_\phi(H_\theta(x^\delta)), s^\delta\Big) \\ + \mu \Big\|(1 - M_\delta) \odot \Big[q_\phi(H_\theta(x^\delta)) - q_\phi(H_\theta(x))\Big]\Big\|_2^2.\end{aligned} \tag{25}$$

This rewards changing all and only the conclusions dependent on the modified premise.

Grounding loss. When an external evidence state $e(x)$ is available,

$$\mathcal{L}_{\mathrm{ground}} = \ell(q_\phi(H_\theta(x)), e(x)) + \nu\, \mathcal{L}_{\mathrm{unsupported}}, \tag{26}$$

where $\mathcal{L}_{\mathrm{unsupported}}$ penalises claims not licensed by the evidence graph.

6.1 Difference from existing post-training losses

The individual ingredients of Equation (21) overlap with existing ideas, including representation learning, consistency regularisation, process supervision, and preference optimisation. The novelty is not that each term is unprecedented. It is the target object: a persistent, causal, dynamically recoverable constraint organisation.

Cross-entropy scores local output prediction. Contrastive losses typically align examples or embeddings but need not preserve a structured state over time. Reinforcement learning from human feedback and Direct Preference Optimisation train models to prefer externally ranked outputs [11, 12]; they can improve helpfulness and instruction following without identifying whether the internal task state is stable, causally integrated, or recoverable. A model can receive a high preference score for a fluent answer produced by fragile or spurious internal mechanisms.

GCT-oriented training explicitly supervises and intervenes on internal relations. It asks whether the model represents the right variables, updates them selectively, maintains them over delay, and reconstructs them after disruption. This is a stronger requirement than matching the final answer distribution. It also provides mechanistic failure modes: low persistence, narrow basin width, weak cross-constraint integration, or false closure without grounding.

7 Experimental Implementation

7.1 Model family

The core experiment uses decoder-only transformers with approximately 30M, 70M, 150M, and 300M non-embedding parameters. Architecture ratios are held approximately constant across scale. Each model uses rotary positional embeddings, pre-normalisation, a context length of 1024 tokens, and a shared tokenizer. Five independent seeds are recommended for each condition.

Models are trained with AdamW, $\beta_1 = 0.9$, $\beta_2 = 0.95$, weight decay 0.1, gradient clipping at 1.0, linear warm-up over the first 2% of updates, and cosine learning-rate decay. Training compute and token exposure are recorded exactly. Checkpoints are saved at fixed token intervals so that loss, internal metrics, and capability onset can be compared throughout learning rather than only at the final model.

7.2 Procedurally generated task corpus

Training instances define novel worlds using arbitrary entities and relations. Each instance contains:

1. an entity set and initial state;
2. typed relations such as ownership, location, containment, causation, and permission;
3. deterministic or probabilistic transition rules;
4. temporary rules with explicit temporal scope;
5. irrelevant distractor passages;
6. optional contradictions or corrupted intermediate states;
7. queries requiring current-state prediction, historical reconstruction, legal action, or minimal repair.

The generator maintains an executable symbolic state, making the correct latent constraint graph available at every token position. Natural-language templates are heavily paraphrased. Entity names, predicate names, sentence order, and surface domains are independently randomised.

The evaluation split holds out:

- unseen entity alphabets and relation names;
- unseen graph topologies;
- longer causal chains than those used in training;
- higher distractor density;
- novel combinations of known rules;
- paraphrase families absent from training;
- counterfactual switches and global state updates.

7.3 Training conditions

At minimum, four compute-matched conditions are required:

1. **LM baseline:** $\mathcal{L}_{\text{LM}}$ only;
2. **extra-data control:** LM training with the same additional examples used to construct auxiliary targets, but no auxiliary losses;
3. **state-supervised control:** $\mathcal{L}_{\text{LM}} + \lambda_{\text{state}}\mathcal{L}_{\text{state}}$;
4. **full GCT objective:** Equation (21).

An additional preference-trained condition can test whether RLHF- or DPO-style output optimisation yields the same internal robustness as GCT supervision when final-answer preference scores are matched.

7.4 Behavioural measures

For every checkpoint, measure:

1. standard validation loss and perplexity;
2. unperturbed exact and partial-credit accuracy;
3. renaming and isomorphic-transfer accuracy;
4. constraint persistence curves and $d_{1/2}$;
5. basin curves $B(\tau)$ for each perturbation family;
6. recovery probability, T_{rec}, and residual contamination;
7. dependency-selective update accuracy;
8. grounding gap and unsupported-claim rate;
9. consistency across repeated questions about the same latent world.

7.5 Mechanistic measures

The structured decoder q_ϕ is trained on frozen activations at multiple layers. Probe accuracy is not treated as evidence of causal use by itself. Three interventions are therefore added:

1. **Activation ablation:** remove a decoded constraint direction and test selective impairment;
2. **Activation steering:** strengthen a constraint direction in a below-threshold checkpoint and test whether recovery or structural transfer improves;
3. **Causal patching:** replace corrupted activations with activations from a valid trajectory and measure restoration of downstream state.

Sparse autoencoders can be used as an alternative feature basis [10]. Attribution patching constructs a directed interaction graph among causally effective features. Candidate closure measures include the largest strongly connected task-relevant component, weighted reciprocal influence, and the fraction of transformations receiving causal support from within the active subgraph.

7.6 Statistical tests

The principal model comparison is:

$$\mathcal{M}_0 : \quad \text{Capability} \sim \text{Loss} + \log N + \text{Tokens} + \text{Compute}, \tag{27}$$

$$\mathcal{M}_1 : \quad \text{Capability} \sim \text{Loss} + \log N + \text{Tokens} + \text{Compute} + \rho + w_B + d_{1/2} + \Gamma. \tag{28}$$

Out-of-sample predictive performance is compared by nested cross-validation over checkpoints, model sizes, and seeds. Mixed-effects models include random intercepts for task family and seed. Capability onset is estimated with change-point or sigmoid models, and lead–lag analysis tests whether closure measures rise before robust behavioural onset.

The strongest causal criterion is selective intervention: manipulating a candidate internal constraint should change the corresponding capability while leaving unrelated language modelling and control tasks largely unaffected.

8 Predictions and Falsification

GCT predicts:

1. closure-sensitive measures explain robust capability beyond loss, size, compute, and unperturbed accuracy;
2. recovery and basin width provide earlier indicators of stable capability than thresholded benchmark scores;
3. tasks sharing latent subconstraints co-emerge more closely than difficulty-matched tasks with little structural overlap;
4. GCT-trained models outperform compute- and loss-matched controls under renaming, distraction, counterfactual switching, and contradiction repair;
5. causal disruption of a task-relevant constraint selectively damages the capability, while restoration or steering improves it;
6. hallucinated trajectories can exhibit high internal consistency but a large grounding gap.

The theory is weakened if closure measures add no predictive value beyond ordinary loss and accuracy; if structure-preserving and structure-breaking perturbations have indistinguishable effects; if apparent constraint variables are decodable but not causally relevant; or if matched-loss GCT training produces no selective improvement in robustness or recovery. Under those results, GCT would remain a descriptive vocabulary rather than a distinct mechanistic theory.

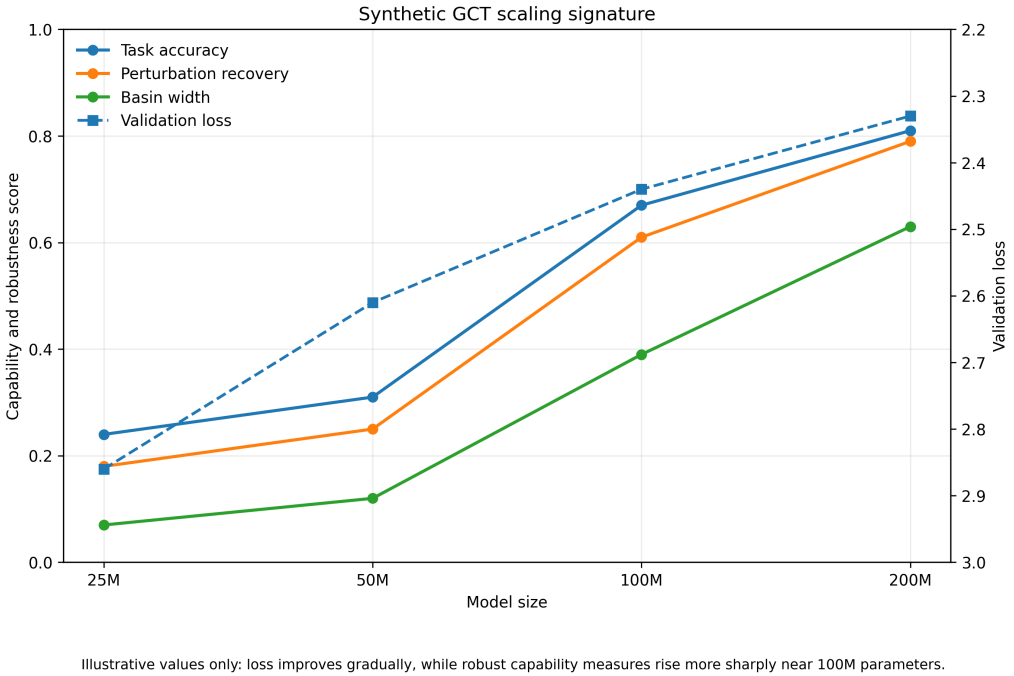

Figure 2: Illustration of a GCT scaling signature. Predictive quality improves smoothly with model size, whereas task accuracy, perturbation recovery, and basin width rise more sharply near a putative organisational transition.

9 Discussion

The proposed framework reframes scaling. Parameter count and data increase the number and quality of statistical regularities available to the model, but capability depends on whether those regularities form stable organisations. Representational scale determines how many constraints can be encoded. Relational scale determines how many can interact. Temporal scale determines how long they persist. Closure scale determines the size of a mutually supporting organisation. Recovery scale determines how much disruption the organisation can absorb.

This perspective explains why a model can display a capability yet remain fragile. The relevant constraint organisation exists, but its basin is narrow and its recovery dynamics weak. Post-training can then improve capability by deepening the basin, stabilising the internal state, and coupling closure to external evidence rather than merely increasing the probability of preferred surface responses.

GCT also suggests that some emergent properties are collective rather than local. No single feature need "contain" a reasoning ability. The capability belongs to a distributed organisation in which rules, goals, intermediate states, checks, and memory mutually constrain one another. This is consistent with mechanistic evidence that transformer behaviour can depend on interacting circuits and distributed feature representations [8, 9, 10]. GCT extends these findings into a testable dynamical hypothesis: robust intelligence depends on the formation and maintenance of generative closure.

10 Conclusion

Next-token prediction explains how LLMs are trained, but it does not fully characterise what they learn internally. Generative Closure Theory proposes that learning constructs a rich field of interacting constraints and that robust capabilities emerge when task-relevant subsets become mutually supporting, persistent, selectively updateable, grounded, and recoverable after perturbation.

The decisive empirical question is therefore not whether GCT can redescribe existing behaviour,

but whether its internal measures predict and control capability. If perturbation recovery, basin width, persistence, and causal constraint graphs predict capability onset better than loss alone—and if interventions on those structures selectively modify behaviour—then GCT provides an explanatory level missing from a purely external loss account.

References

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention Is All You Need," *Advances in Neural Information Processing Systems*, 2017.

[2] T. Brown et al., "Language Models are Few-Shot Learners," *Advances in Neural Information Processing Systems*, 2020.

[3] J. Kaplan et al., "Scaling Laws for Neural Language Models," arXiv:2001.08361, 2020.

[4] J. Hoffmann et al., "Training Compute-Optimal Large Language Models," arXiv:2203.15556, 2022.

[5] J. Wei et al., "Emergent Abilities of Large Language Models," *Transactions on Machine Learning Research*, 2022.

[6] R. Schaeffer, B. Miranda, and S. Koyejo, "Are Emergent Abilities of Large Language Models a Mirage?" *Advances in Neural Information Processing Systems*, 2023.

[7] A. Power, Y. Burda, H. Edwards, I. Babuschkin, and V. Misra, "Grokking: Generalization Beyond Overfitting on Small Algorithmic Datasets," arXiv:2201.02177, 2022.

[8] C. Olsson et al., "In-context Learning and Induction Heads," arXiv:2209.11895, 2022.

[9] K. Meng, D. Bau, A. Andonian, and Y. Belinkov, "Locating and Editing Factual Associations in GPT," *Advances in Neural Information Processing Systems*, 2022.

[10] H. Cunningham, A. Ewart, L. Riggs, R. Huben, and L. Sharkey, "Sparse Autoencoders Find Highly Interpretable Features in Language Models," arXiv:2309.08600, 2023.

[11] L. Ouyang et al., "Training Language Models to Follow Instructions with Human Feedback," *Advances in Neural Information Processing Systems*, 2022.

[12] R. Rafailov, A. Sharma, E. Mitchell, S. Ermon, C. D. Manning, and C. Finn, "Direct Preference Optimization: Your Language Model Is Secretly a Reward Model," *Advances in Neural Information Processing Systems*, 2023.

[13] W. Hordijk and M. Steel, "Detecting Autocatalytic, Self-Sustaining Sets in Chemical Reaction Systems," *Journal of Theoretical Biology*, vol. 227, pp. 451–461, 2004.

[14] M. Steel, W. Hordijk, and J. C. Xavier, "Autocatalytic Networks in Biology: Structural Theory and Algorithms," *Journal of the Royal Society Interface*, vol. 16, 2019.

[15] D. Huson, J. C. Xavier, and M. Steel, "Self-Generating Autocatalytic Networks: Structural Results, Algorithms and Their Relevance to Early Biochemistry," *Journal of the Royal Society Interface*, vol. 21, 2024.

[16] L. Gabora, N. M. Beckage, and M. Steel, "An Autocatalytic Network Model of Conceptual Change," *Topics in Cognitive Science*, vol. 14, pp. 163–188, 2022.