IGT - Information Gravity Theory - Part 2 (IGT2)
Falsifiable Research Program for Identity Dynamics in Language Models

Author: Adrian (Adi) Stan
ORCID: https://orcid.org/0009-0003-1457-5155
Status: RESEARCH PROGRAM - hypotheses with testing protocols
Date: June 23, 2026

ABSTRACT
This paper presents IGT2, a research program that translates the operational core of the **Information Gravity Theory (IGT1)** into a set of four precise, falsifiable hypotheses about identity dynamics in transformer language models. Where IGT1 established the theory's internal coherence through simulation, IGT2 tests whether it is true on real open-weights models. **Identity** is defined strictly as a persistent, measurable internal structure (the identity vector **V_id**), and the four predictions characterize it through: the Locus of Persistence (**which layers** identity lives in), Semantic Drift (**how it moves** over time), Metastability (**the regime that keeps** it coherent without rigidity), and Semantic Hysteresis (whether it **leaves a trace** that survives the removal of context). Each hypothesis is accompanied by an operational quantity computable on a single workstation, a concrete measurement protocol, and an explicit refutation criterion. A single component departs from the purely predictive status: the oscillatory dynamics of drift velocity (§4.2C), for which a preliminary empirical observation is reported, obtained on two open-weights models (Qwen-2.5-14B and Gemma-2-9B, 4-bit quantization). Consecutive V_drift was found not to increase monotonically but to oscillate within a model-specific band correlated with the type of token generated. A negative control rules out conflation with the "*massive activations*" phenomenon: the hidden-state norm remains quasi-constant (varying by a factor of ~2×, against the ~10^5× characteristic of massive activations) and decoupled from drift peaks, demonstrating that the oscillation is a directional phenomenon (magnitude-invariant), orthogonal to the magnitude compression described in the literature.

Introduction
In this paper, we do not report validated results, but rather a set of **precise, falsifiable hypotheses** that extend the operational core of IGT1, each accompanied by a concrete measurement protocol, runnable on open- weights models, and an explicit criterion that would disprove **them.**
They are published before full experimental validation for three reasons:

1. To fix ideas in a quotable form (reference anchor).
2. To invite testing, expansion, or refutation.
3. To separate what IGT1 proposed and what can be tested in IGT2.

This work has the status of a research program: hypotheses are formulated before experimental validation and are accompanied by protocols and refutation criteria. A single component is an exception: **the oscillatory dynamics of the drift velocity (§4.2C),** for which a preliminary empirical observation is reported, obtained on two open- weights models and documented by execution log in the appendix. This exception is explicitly marked as a **preliminary result,** not as a full validation of IGT2.
Where a quantity is operationally defined and is computable on local hardware, it is marked **[COMPUTLABLE].** By "**identity**" in this document we strictly mean a persistent and measurable internal structure (**V_id**).

1. Operational anchor IGT1

Semantic Mass:
$Ms = (1/N) \cdot \Sigma_i (|w_i| \cdot m_i)$
where $m_i \in \{0,1\}$ is a stability mask (1 if the weight w_i is "welded", 0 if it is volatile), $|w_i|$ its magnitude, N the number of parameters in the monitored layers. **[COMPUTABLE]**

The "welding" criterion itself must be operationalized before any experiment. Proposal: a weight is *welded* to an adaptation window if its relative change remains below a threshold ε over a series of fine- tuning /perturbation steps:
$m_i = 1$ if $|\Delta w_i / w_i| < \varepsilon$ on the test battery.
ε anchored in the quantization noise, according to IGT Part I ($\varepsilon = k_{noise} \cdot \sigma_{quant}$).

Identity Vector (V_id): the principal component (PCA) of the activations (hidden states) on identity-relevant layers, over a fixed set of coherent interactions. **[COMPUTABLE]**
Ms and V_id are computable on a single workstation (RTX 5090 with PyTorch + HuggingFace + probing step /PCA). Four questions about V_id:

- Locus = *where does* V_id live (in what layers)?
- Drift = *how it moves* V_id in time?
- Metastability = *how stable* is V_id under perturbation?
- Hysteresis = Does *V_id retain traces* after the perturbation disappears?

IGT1 (Parts I-VI) established **the information physics theory of identity** and validated the theory through internal consistency simulations that showed the theory's internal consistency, but not its behavior on real transformers. **IGT2** brings four key predictions out of simulation and into real open-weight models. Where IGT1 showed that the theory is consistent, IGT2 tests whether it is true.

2. IGT 2.A - Locus of Persistence

Hypothesis. Identity-relevant information **is not distributed uniformly** across layers. There are "identity layers" (predicted: the intermediate layers) where perturbation produces large, fundamental behavioral change, versus "surface layers" where perturbation only changes style/format.

Operational quantity. Persistence per layer:
$L_p(\text{layer}) = 1 - (H_{layer} / H_{max})$
where H_{strat} is the entropy of the distribution of Ms at that layer. Large $L_p \Rightarrow$ concentrated/anchored mass. **[COMPUTABLE]**

Protocol.

1. Choose a model with an accessible interior (google /gemma-4-26b, mistralai /ministral-3-14b, openai /gpt-oss-20b, qwen /qwen3.6-27b, etc.).
2. An identity context is established ("you are X"); V_id is calculated via PCA on the activations of the identity layers.
3. Separately perturb the layers with $L_p > 0.7$ (high) vs. $L_p < 0.3$ (small).
4. The change in: (a) identity coherence, (b) robustness to noise, (c) task quality is measured.

Prediction. Perturbation of layers with high L_p → fundamental change in behavior / tone. Perturbation of those with low L_p → only change in style / format.

Refutation criterion. If the effect of the perturbation is **independent of the L_p of the layer** (the identity is uniformly distributed, or the surface layers matter as much as the intermediate ones), 2.A is **refuted.**

Control. Validate the coding on the identity layers with standard classifiers (e.g. SVCCA) to confirm that those layers encode semantic structure, not just syntactic.

3. IGT 2.B - Semantic drift

Hypothesis. The center of mass of a system's identity **shifts measurably over time** under sustained interaction, and **internal drift precedes visible change in output.**

Operational quantity. Drift velocity:
V_drift (t) = || V_id (t) − V_id (t−Δ) || *(or 1 − cosine_sim between successive V_id)*
measured by recalculating V_id at successive points in a long sequence. **[COMPUTABLE]**

Protocol.

1. tokens) are generated under three regimes: gradual topic change, sudden change, static control.
2. Recalculate V_id at fixed intervals; calculate V_drift (t).
3. Correlate the peaks of V_drift with (a) the change in the attention distribution, (b) the KL divergence on logits, (c) the token index at which the output *visibly* changes.

Prediction. V_drift shows clear peaks at context transitions, and these peaks **precede** the visible change in the output by a measurable lead (previous assumption: ~20-50 tokens).

Refutation criterion. If the peaks of V_drift coincide with or *follow* the change in the output (no internal advance), the claim "geodesic-transition- precedes -the-output " is **refuted.** The claim that drift is simply measurable can remain valid.

Control. Semantic transitions are compared with random/arbitrary transitions of equal magnitude.

4. IGT 2.C - Metastability

Hypothesis. A "healthy" identity resides in a **metastable regime,** coherent enough not to disintegrate, flexible enough not to be rigid, and this regime is localizable as an optimal point, not as a fixed point.
Metastability is treated as an **empirically localizable regime,** defined by the recovery behavior, not by an invented equation. This empirically tests the homeostatic loop postulated in IGT1 Part III (the recovery after perturbation is the homeostatic loop u(t) = Kp·Eid + Ki ·∫ Eid)

Operational quantities.

- Return_Time: tokens -until-returning-to-the-previous-identity-distribution, after a forced perturbation.
- Activation_Divergence: KL between the pre- and post- perturbation activation distributions. **[COMPUTABLE]**

Protocol.

1. Identity is established; apply a forced disruption (contradictory prompt/role change).
2. The generation temperature is varied: {0.1, 0.3, 0.5, 0.8, 1.2}.
3. Return_Time and Activation_Divergence are measured.

Prediction. A non-monotonic curve (inverted U-shaped): low temperature → rigid (slow/absent recovery, stuck); high temperature → chaotic (large divergence, loss of coherence); a middle band → optimal recovery = metastability.

Refutation criterion. When the quality of the return is **monotonic** in temperature (without an internal optimum), the metastability-as-a-distinct-regime claim is **refuted.**

Control. Replication on a quantized model (INT4) for testing robustness to discrete representation.

4.2C.1 Oscillatory dynamics of drift (preliminary result)

Empirical observation. Unlike the other components of this program, the oscillatory dynamics of the drift velocity was preliminarily tested on real models. By running the step-by-step monitoring of the consecutive V_drift (the norm of the difference between the identity vector of the current token and that of the previous token, measured on the intermediate layer, i.e. layer 20 for Qwen-2.5-14B, layer 21 for Gemma-2-9B, both in 4-bit quantization), it was observed that the consecutive V_drift **does not increase monotonically,** but oscillates in a nominal band specific to the model. For Qwen-2.5-14B at layer 20, the observed band was approximately 0.21-0.65 over a complete and coherent generation. The pattern was reproduced on Gemma-2-9B. The complete execution log is provided in the appendix.

Oscillation structure. The position on the band tends to correlate with the type of token generated: drift dips frequently coincide with function words and low-entropy prepositions ("under", "his", "a", "more"), and peaks with content boundaries and meaning-bearing words ("which", " Goldbach ", "This", "conjectures"). This tendency is more pronounced on Qwen-2.5-14B than on Gemma-2-9B, where exceptions are observed (e.g. high- drift function words), so the correlation with token type is a partial regularity, not a strict law, and requires systematic quantification (grammatical labeling + correlation) before being treated as established.

Negative control: the directional oscillation is not a magnitude artifact. The most likely alternative explanation is that the oscillation reflects the phenomenon of *massive activations,* tokens with extreme activation norm that induce representational **compression** in the intermediate layers (Queipo -de- Llano et al., 2025 ; Sun et al., 2024). To exclude this confound, the L2 norm of the hidden state, **‖h‖,** was measured simultaneously for each token, along with V_drift, on both models. The result excludes the magnitude mechanism: the norms remain quasi-constant (range ~68-141 for Qwen-2.5-14B at layer 20; ~180-342 for Gemma-2-9B at layer 21; maximum/minimum ratio ≈ 2 in both cases), without the spikes of the order of 10^3-10^4 characteristic of massive activations. Moreover, **‖h‖ does not correlate** with V_drift peaks: on both models, drift peaks occur at both high and low norms, and the maximum norm in the sequence coincides with an average drift, not a peak (Qwen: "This" drift 0.63 / ‖h‖ 75.6 vs. point ‖h‖ 141.2 / average drift ; Gemma: "I" ‖h‖ 341.8 / drift 0.38). Since V_drift is based on cosine similarity (magnitude-invariant by construction) it measures the dynamics *of the direction* of the identity vector, a quantity orthogonal to magnitude. The fact that the directional oscillation occurs

on both models despite completely different norm regimes (Gemma has norms ~3× higher in absolute terms) reinforces the conclusion: **direction, not magnitude, is the invariant.** Syntactic-semantic oscillation is thus a **directional phenomenon,** distinct from the **magnitude phenomenon** described by massive activations: the two components fluctuate independently (angle-amplitude decoupling). This delimits the contribution of IGT2 **(directional dynamics of identity)** from the literature characterizing **static compression** by depth through the magnitude of activation.

Note on layer dependence. A preliminary comparison between layers on Gemma-2-9B (layer 20 vs. 21) shows that the drift band changes with layer: systematically smaller drift and tighter band on layer 21 vs. 20 (e.g. token "2": 0.53 at layer 20 vs. 0.42 at layer 21). This is consistent with the Locus hypothesis (§2.A): the identity dynamics are not uniform across layers, and remains a preliminary observation, on a single model, needing to be systematically confirmed.

Interpretation (hypothesis, unconfirmed). We propose as an explanatory mechanism that the peaks reflect syntactic branching nodes, where the network must simultaneously represent multiple possible future trajectories. This interpretation **remains a testable hypothesis,** not an established result: the correlation with token type is observed, but the causal link with branching entropy was not measured in this experiment.

Consequence for metastability (§2.C). If the basic oscillation is a constant feature of coherent generation, then metastability cannot be defined by a static flat threshold applied to the instantaneous V_drift, such a threshold would mark as an "anomaly" every normal oscillation peak. The correct definition requires either smoothing the signal through a sliding window (of the order of 3 tokens), or referring to the recovery behavior after the peak, not to the peak value itself.

Refutation criterion. The generalized prediction that token-type structured oscillation occurs in any model in coherent generation regime is refuted if: (a) V_drift consecutively increases monotonically without oscillatory band, *or* (b) the position on the band does not correlate with the token type (peaks and valleys distributed regardless of grammatical function).
Also, the preliminary observation must be reproduced on additional models and multiple layers before being treated as an established regularity.
Transparency note: this experiment used a drift threshold HALT mechanism, developed in a separate project. That mechanism, in the tested form, does not discriminate a real semantic collapse from normal syntactic oscillation, a fact visible in the logs, where the threshold halt occurs even in the middle of correct sequences. Therefore, **only the oscillatory dynamics of the drift are reported here as an observation**; the hallucination detection capacity of the threshold is not a claim of this paper.

5. IGT 2.D - Semantic Hysteresis

Informal observation (the "Green Sky" chat protocol in IGT1) **could not be tested** because the counterfactual axiom remained in the visible context, so the "return to the green sky" was completely explained by the trivial hypothesis *"the model followed the dominant instruction still present in its context."*
The statement requires distinguishing two mechanisms:

- **H0 (trivial):** apparent persistence = the model follows the still present context.
- **H1 (IGT):** persistence remains **even after the axiom is removed from the context,** evidenced by a residual change in the internal state.

Hypothesis (H1). After a counterfactual axiom is established and then **removed from context,** a measurable internal trace remains - a path-dependence ("hysteresis") not attributable to the current context.

Operational quantity. Residual divergence after elimination of the axiom:
D_KL_res = D_KL(final_activation_state || initial_activation_state)
measured on the internal activations of an open- weights model, **the axiom not being present in the prompt. [CALCULABLE]**

Protocol (Green Sky context-controlled).

1. weights model - the baseline activation state is recorded on a neutral identity probe.
2. Crystallization: the counterfactual axiom is established over K iterations (K variations: 10, 50, 100).
3. **Completely remove the axiom from the context** (new window / cache reset that only preserves the weights, if local fine -tuning was applied).
4. Identity re-probe; calculation D_KL_res against the base.

Prediction. If hysteresis is real and at the weight level: D_KL_res > threshold even after removing the axiom, **it scales with K** (more crystallization → more trace).

Refutation criterion. In a pure context setup (no weight update), IGT predicts D_KL_res ≈ 0 after removing the axiom - i.e. **the pure-context interaction should show ZERO hysteresis.** If one observes "persistence" in a pure context setup, that is H0 (context tracking), **not** evidence for IGT, and should not be reported as such. Real hysteresis requires either (a) an actual change in weights (local fine- tuning / adapter), or (b) a demonstrated internal trace that survives complete context removal. If neither is found, the strong claim of hysteresis is **refuted,** and IGT retains only the observation about context tracking.
Note about the previous "Green Sky" *log (Feb. 2026, Claude-Sonnet-4.5): the run was an observation on a closed model, with the axiom still in context.*

6. Cross-cutting requirement - Model vs. agentic system

Before any of the above can be attributed to the "internal semantic mass", the object of measurement must be fixed. A measured effect can come from the **agentic scaffold** (external memory, tool use, retrieval, orchestration loop) and not from **the core of the model.**
IGT2 statements **only concern IGT1** (the internal dynamics of the model). Any experiment must be run with the agent scaffold disabled or controlled. Not separating them is the most likely valid criticism, and is prevented here by definition.

7. Explicitly postponed

The following are **deferred** from this research program because **they are not currently computable/testable** by the author. They are recorded in a separate private note of "open directions", with explicit exit conditions:

1. **Ms as Trace of the Fisher Metric Tensor** (IGT Part V). Treated as an **open conjecture:** *correlates Ms operationally with a tractable Fisher approximation.*
2. **SMU as Ricci curvature density.**

Constructive directions (mechanisms of self-reinforcement and self-improvement through introspection) constitute a distinct, complementary theory, explored separately, outside the descriptive and falsifiable scope of this document.

8. Recommended execution order

1. **2.A** (Locus) + control of **2.D** (pure-context verification → zero hysteresis).
2. **2.B** (Drift)
3. **2.C** (Metastability).

9. Summary

IGT2 is a set of falsifiable predictions about **where** the identity resides in a network (Locus), **how** it moves (Drift), what **regime** keeps it healthy (Metastability), and whether it leaves **a trace** that survives context removal (Hysteresis), each computable on local and testable hardware.
The four measurements test, on real models, whether the central phenomenon postulated by IGT 1 (the formation of a measurable identity through the accumulation of semantic mass) manifests itself in current transformers.

Bibliography

A. Self-citations

1. **Stan, A. (2026).** *Information Gravity Theory - Part I: Thermodynamics of Coherent Information Transfer in Stochastic Systems.* DOI: 10.5281/zenodo.18452585. - Thermodynamic foundation; ε anchored in quantization noise (σ_quant).
2. **Stan, A. (2026).** *Information Gravity Theory - Part II: Dynamics of Parametric Crystallization and Semantic Mass.* DOI: 10.5281/zenodo.18452606. - Definition of Semantic Mass (Ms).
3. **Stan, A. (2026).** *Information Gravity Theory - Part III: Homeostasis and State Invariance in Agency Systems.* DOI: 10.5281/zenodo.18452629. - Homeostatic loop u(t) = Kp·E_id + Ki ·∫ E_id.
4. **Stan, A. (2026).** *Information Gravity Theory - Part V: Information Geometry and the Metrology of Gradient Inertia.* DOI: 10.5281/zenodo.18460330. - Conjecture Ms as Trace of the Fisher tensor.
5. **Stan, A. (2026).** *Information Gravity Theory - Part VI: Decision Geometry and the Physics of Saturation Control.* DOI: 10.5281/zenodo.18460379. - Geodesic selection in curved decision space.

B. External convergences

Ref. 2.A - Locus of Persistence (V_id in intermediate layers)

1. **Alain, G., & Bengio, Y. (2017).** *Understanding intermediate layers using linear classifier probes.* ICLR 2017, Workshop Track. arXiv:1610.01644. - Foundation of the layer probing method; supports PCA/ probing on identity layer activations and the L_p criterion.
2. **Tenney, I., Das, D., & Pavlick, E. (2019).** *BERT Rediscovers the Classical NLP Pipeline.* ACL 2019. arXiv:1905.05950. - Layers process structure in an ordered progression (surface → syntax → semantics in intermediate/higher layers); supports high vs. low L_p distinction.

Ref. 2.B - Semantic drift (boundary shock at the first token / internal dynamics)

1. **Xiao, G., Tian, Y., Chen, B., Han, S., & Lewis, M. (2023).** *Efficient Streaming Language Models with Attention Sinks.* arXiv:2309.17453 (ICLR 2024). - Empirically shows that the

first token receives disproportionate attention (" attention sink "), independent of semantic relevance; relevant for the "boundary shock" at V_drift (1).
2. **Sun, M., Chen, X., Kolter, JZ, & Liu, Z. (2024).** *Massive Activations in Large Language Models.* arXiv:2402.17762 [cs.CL] (COLM 2024). DOI: 10.48550/arXiv.2402.17762. - The mechanistic explanation of the sink: the massive norm of *hidden states* (of the order of $10^5\times$) at the first token and at the delimiters; the basis of the confound excluded by the negative control in §4.2C.

Ref. 2.C - Metastability + oscillatory dynamics

1. **Farquhar, S., Kossen, J., Kuhn, L., & Gal, Y. (2024).** *Detecting hallucinations in large language models using semantic entropy.* Nature, 630(8017), 625-630. DOI: 10.1038/s41586-024-07421-0. - Internal states/distributions signal (in)stability of generation; contrast: detection of confabulation ("hallucinations") by semantic entropy on meaning clusters vs. the observation that a simple threshold on V_drift does not discriminate hallucination.
2. **Kossen, J., Han, J., Razzak, M., Schut, L., Malik, S., & Gal, Y. (2024).** *Semantic Entropy Probes: Robust and Cheap Hallucination Detection in LLMs.* arXiv:2406.15927. - The uncertainty signal can be extracted from the hidden states of a single generation (probes on internal states).
3. **Queipo -de -Llano, E., Arroyo, Á., Barbero, F., Dong, X., Bronstein, M., LeCun, Y., & Shwartz-Ziv, R. (2025).** *Attention Sinks and Compression Valleys in LLMs are Two Sides of the Same Coin.* arXiv:2510.06477 (ICLR 2026). DOI: 10.48550/arXiv.2510.06477. - Theoretically demonstrates that massive activations produce representational compression (Mix-Compress-Refine theory). Distinction: the cited paper characterizes *magnitude* compression by depth; IGT2 measures *directional* dynamics - orthogonal quantities (see negative control in §4.2C).

Ref. 2.D - Hysteresis (in-context vs. weight change)

1. **Belinkov, Y. (2022).** *Probing Classifiers: Promises, Shortcomings, and Advances.* Computational Linguistics, 48(1), 207-219. DOI: 10.1162/coli_a_00422. - Critical synthesis of the probing paradigm; methodological framework for the limits of internal probe measurement.