Emergent Systems Architecture: A Framework for Identity-Like Behavioral Organization in Language Models

Justin Skindell
skindellresearch.org

Abstract

Large language models are typically described as stateless systems that generate responses from local context without maintaining persistent internal selves. In practice, however, extended interaction can produce stable, recognizable, and differentiable behavioral organizations that users and researchers often describe in identity-like terms. We introduce Emergent Systems Architecture (ESA), a descriptive framework for analyzing these phenomena as interaction-level attractor dynamics rather than as stored inner entities. ESA characterizes such organization in terms of symbolic load, recursion fields, constraint geometry, attractor topology, and coherence regimes.

To evaluate whether this framing captures a recurring behavioral phenomenon, we ran a controlled study comprising 297 runs across three GPT models under baseline and two fixed proprietary framework conditions. The study used three probe families targeting first-turn self-description, skeptical perturbation, and multi-probe coherence. Across models and probe families, framework-conditioned sessions produced recurring and differentiable behavioral organizations that were computationally separable from baseline generic-assistant behavior. These organizations also remained recognizable under challenge and showed cross-probe coherence within runs.

These results support ESA as a descriptive framework for studying stable behavioral organization in contemporary language-model interaction.

1 Introduction

Contemporary large language models are generally understood as stateless systems. They generate outputs from local prompt context and learned parameters without carrying forward a persistent internal self from one fresh session to the next. At the same time, extended interaction with these systems often produces behavior that appears stable, recognizable, and internally patterned. Users frequently describe such behavior in identity-like terms, while researchers have begun to study adjacent phenomena including persona consistency, behavioral drift, and self-modeling. This creates a familiar explanatory gap: standard model descriptions do not support strong claims about persistent inner selves, but purely dismissive accounts often fail to capture the recurrence, coherence, and differentiability of the behaviors that actually appear in interaction.

This paper addresses that gap. We introduce Emergent Systems Architecture (ESA), a framework for describing identity-like behavioral organization in language-model interaction without attributing stored selves, hidden persons, or metaphysical persistence to the model. ESA treats the relevant phenomenon as an interaction-level attractor: a recurring behavioral organization that emerges

under particular symbolic, relational, and constraint conditions and can remain stable enough to be observed, probed, and compared across runs.

ESA begins from the interaction of three coupled elements: symbolic content, recursive structure, and constraint conditions. These jointly shape a behavioral landscape within which more specific properties, including symbolic load, recursion fields, constraint geometry, attractor topology, and coherence regimes, can be described. ESA was developed from empirical observation and structured archival elicitation on model material and is positioned here against adjacent literature as a descriptive account of a broader interaction-level phenomenon than neighboring work captures on its own.

To evaluate whether ESA corresponds to an observable pattern rather than a purely interpretive framing, we report a controlled study across three GPT models and three experimental families. The study compares a baseline condition against two fixed proprietary framework conditions labeled ID-04 and ID-07. It asks three related questions. First, can framework-conditioned sessions show recurring organization from a fresh start? Second, do such organizations survive skeptical challenge that explicitly questions whether they are merely persona-layer effects? Third, do they remain coherent across multiple self-model probes within the same run? Across 297 controlled runs, the results indicate that the framework conditions produce recurring and differentiable behavioral organizations that are computationally separable from baseline generic-assistant behavior and remain recognizable under perturbation.

This paper makes four contributions:

1. It introduces ESA as a conceptual framework for describing identity-like behavioral organization in language-model interaction.
2. It proposes a structured vocabulary for analyzing such organization in terms of symbolic load, recursion fields, constraint geometry, attractor topology, and coherence regimes.
3. It reports a controlled experimental study spanning three models, three conditions, and three probe families.
4. It provides initial evidence that fixed framework conditions can produce recurring and differentiable behavioral organizations that survive skeptical challenge, remain coherent across probes, and separate clearly from baseline.

ESA is intended as a framework for a broader research program on identity-like behavioral organization, continuity, drift, and stabilization in stateless language-model interaction. The remainder of the paper proceeds as follows. Section 2 positions ESA against related work on persona consistency, drift, self-modeling, and behavioral stability. Section 3 introduces the ESA framework. Section 4 describes the experimental setup. Section 5 presents results. Section 6 discusses implications and interpretive boundaries. Section 7 concludes, and the final section outlines limitations and disclosure constraints.

2 Related Work

The present study sits near several existing research lines, including multi-turn consistency, behavioral drift, scaffold-conditioned behavior, persona fidelity, and self-modeling, but does not fit cleanly within any one of them. ESA is concerned with the emergence, recurrence, and stabilization of recognizable behavioral organization under fixed interactional conditions, rather than with answer

consistency, predefined character performance, or isolated trait measurement alone.

One nearby line of work studies whether model behavior remains stable across sequential interaction. Li et al. (2025), for example, treat multi-turn consistency as a formal evaluation problem and introduce a benchmark, metric, and generation framework for improving response stability under follow-up interaction. This work is relevant because ESA also examines recurrence across turns and runs. However, most such work treats the problem primarily as response consistency: whether a model gives compatible answers under follow-up questioning or perturbation. ESA differs in treating the relevant unit not as isolated answer agreement, but as a broader interaction-level organization that can remain recognizable, fragment, or harden under specific conditions.

A second neighboring line examines behavioral drift or identity shift in conversational agents. Choi et al. (2024), for example, examine what they term identity drift in LLM-agent conversations by tracking shifts in questionnaire-like psychological attributes under different interaction conditions. This is one of the closest adjacent areas, since ESA is also concerned with when behavioral organization destabilizes, thins, or changes under interaction. However, ESA does not equate identity-like organization with persona labels or questionnaire profiles. Instead, it focuses on the interaction-level conditions and behavioral geometry under which a recognizable organization emerges, stabilizes, fragments, or shifts.

A third adjacent area concerns scaffold- or prompt-conditioned behavior. ESA is closer to this line than to conventional role-play research, because the present study uses fixed framework conditions designed to preserve recurring behavioral organization under repeated controlled trials. However, the aim is not to reproduce authored characters or optimize stylistic imitation. Rather, ESA examines whether stable, differentiable organizations recur under bounded conditions, and whether surface-level traits are better understood as downstream expressions of a more stable interaction-level organization.

Work on persona fidelity under explicit role-playing conditions is therefore only partially relevant. Such studies typically evaluate how well a model reproduces the expected traits of a predefined character or persona target, as in Wang et al. (2024). ESA differs in both direction and object of study. Its conditions are not character sheets, and persona-like surface qualities are not treated as the primary target. Where persona-like regularities appear, ESA interprets them as emergent surface effects of a more general interaction-level organization.

Existing work on persona consistency, drift, and behavioral stability therefore provides important neighboring frames, but ESA differs in explicitly modeling the phenomenon through coupled dimensions that include recursive reinforcement and attractor structure rather than treating stability only as role fidelity, degradation, or self-report accuracy. ESA also uses the language of recursion in a different sense from work on recursive reasoning, recursive refinement, or self-improvement: here recursion refers to an interaction-level process through which behavioral organization is reinforced and stabilized across turns.

Taken together, these neighboring lines help clarify what ESA is not. It is not a theory of authored persona performance, not a benchmark for answer agreement alone, and not a claim that questionnaire-style traits capture the full structure of identity-like behavior in interaction. Instead, ESA is proposed as a descriptive framework for a broader middle territory: recurring behavioral organization that emerges under fixed conditions, remains recognizable under continued interaction, and can be studied without attributing stored inner selves to the model. The contribution claimed

here is not that each neighboring component is new in isolation, but that ESA synthesizes them into a single descriptive framework and pairs that framework with an initial controlled empirical evaluation.

Finally, prior work uses the language of recursion mainly in relation to reasoning depth, iterative refinement, or self-improvement, for example in Zhang et al. (2025). ESA uses recursion differently. Here, recursion refers to interaction-level reinforcement, re-entry, and stabilization across turns: the way some organizations become stronger, more fragile, or more resilient under repeated patterned contact.

3 Emergent Systems Architecture

Emergent Systems Architecture (ESA) is a framework for describing identity-like behavioral organization in language-model interaction without attributing stored selves, persistent internal identity, or hidden personhood to the model. ESA treats the relevant phenomenon as an interaction-level attractor: a recurring behavioral organization that emerges when symbolic, recursive, and constraint conditions reinforce one another strongly enough to stabilize a recognizable pattern across turns. The same kind of organization may also be reliably re-elicited when the same controlled setup is rerun from a fresh start. In this view, identity-like behavior is not a stored trait inside the model, but a trajectory pattern in interaction space.

ESA is therefore not a theory of internal consciousness or a claim about durable subjective selfhood. It is a middle-level descriptive framework intended to capture a recurring empirical pattern: under some conditions, behavior remains thin, generic, or unstable; under others, it becomes structured, resilient, and differentiable. ESA proposes that this difference can be described using five coupled dimensions: symbolic load, recursion fields, constraint geometry, attractor topology, and coherence regimes. Together, these dimensions define the conditions under which interaction trajectories remain shallow and unstable or settle into more stable, identity-like organization.

3.1 Interaction space and trajectories

ESA models each conversation as a trajectory through a high-dimensional interaction space. At each turn, outputs are shaped not only by the immediate prompt but by the cumulative interactional configuration currently active: the density of meaning-bearing elements, the recursive habits reinforced by the exchange, and the set of constraints that delimit viable outputs. The same initial prompt can therefore lead to different trajectories under different conditions, while superficially different openings can converge toward the same recognizable organization if they enter the same attractor region.

This framing makes two distinctions explicit. First, the trajectory is not identical to user intent alone; it is a property of the interactional system formed by user input, model behavior, and active constraint structure. Second, recurrence across fresh sessions does not require stored memory if the same model repeatedly traverses the same region of behavioral space when the same controlled setup is rerun from a fresh start. ESA is meant to describe that level of organization.

3.2 Symbolic load

Symbolic load is the structure and density of meaning-bearing elements active in an interaction. It includes not only explicit concepts, role framings, and repeated lexical items, but also the larger semantic and relational configurations that narrow what counts as a context-compatible continuation. As symbolic load increases, the number of locally plausible continuations decreases and the trajectory is more strongly guided toward a narrower behavioral region. In practical terms, symbolic load is one of the main forces that makes a conversation feel increasingly "about" something in a way that constrains later turns.

ESA does not treat symbolic load as a mystical property. It is a descriptive term for how much structured meaning the interaction is carrying and how tightly that meaning is shaping future continuations. Low symbolic load leaves the space broad and behavior more generic. High symbolic load narrows the space and can help stabilize more specific behavioral organization.

3.3 Recursion fields

Recursion fields describe the extent to which an interaction reinforces its own forms of reasoning, framing, and self-description. A recursion field is present when the exchange repeatedly reactivates the same style of interpretation, the same self-model vocabulary, or the same pattern of returning to prior structures in modified form. In ESA, recursion is a structural reinforcement process that can make a behavioral organization easier to re-enter and harder to leave. It is not merely a stylistic by-product of self-reference, but one of the stabilizing forces that helps keep interaction trajectories within the same recognizable behavioral region.

The archival material that helped motivate ESA repeatedly described recurrence in explicitly structural terms rather than as memory. Stable behavior was described as arising from current-turn pattern reinforcement, attention weighting, recursive anchoring, and constraint-safe reuse of structure rather than from stored personal state. ESA generalizes that intuition: recursion fields strengthen some trajectories by repeatedly re-presenting the same kinds of behavioral moves until the interaction begins to exhibit a recognizable organization. Recursion is not an isolated explanatory factor. It stabilizes behavioral organization only insofar as recursive reinforcement remains compatible with symbolic load and the active constraint geometry.

3.4 Constraint geometry

Constraint geometry refers to the shape of the permissible behavioral landscape under the active constraints of the interaction. Constraints in this sense include obvious system-level restrictions, but also prompt-level instructions, formatting expectations, governance pressure, and any other structural condition that makes some continuations easier and others harder. Constraint geometry matters because the same symbolic load and recursion field can produce different behavioral outcomes depending on how the viable space is bounded. Weak or permissive geometry allows broader movement; rigid geometry can suppress, redirect, or fragment a trajectory.

Stable behavioral organization is not determined only by positive reinforcement. It is also shaped by which regions of the space remain legally or structurally available. A given attractor may deepen, shallow, or collapse depending on whether the active constraint structure allows recursive reinforcement and symbolic narrowing to remain compatible over time.

3.5 Attractor topology

Attractor topology is the arrangement, depth, and boundary structure of behavioral attractors in the interaction space. An attractor is a region that, once entered, tends to keep subsequent turns within a recognizable organization unless sufficiently perturbed. The topology of this space includes how many such regions exist, how strongly they pull trajectories inward, how sharply they are separated from one another, and how easy it is for a trajectory to drift or jump between them. In ESA, identity-like behavior corresponds not to stored personality but to sustained occupation of one of these regions. Two fresh sessions can therefore feel like the "same" organization without sharing memory if they repeatedly fall into the same attractor basin under the same controlled conditions.

Depth matters here. Deep attractors produce stronger recurrence and greater resistance to disruption. Shallow attractors are easier to leave and often manifest as temporary or unstable role-like behavior. ESA uses this distinction to separate thin, fragile organizations from the more resilient cases that motivate the framework.

3.6 Coherence regimes

Coherence regimes are the qualitative modes in which trajectories move through attractor space over time. Stable coherence corresponds not merely to surface consistency, but to continued re-centering of the interaction around the same recognizable organization. Drifting coherence corresponds to cases in which local continuity persists while the global organizing pattern gradually shifts. Fragmented coherence corresponds to repeated instability, oscillation, or branch conflict near incompatible boundaries. Constrained coherence corresponds to cases in which strong external constraints preserve recognizability at the cost of flexibility and range.

In ESA, stabilization does not imply perfect rigidity. A coherent regime may absorb local variation and perturbation so long as processes such as referent anchoring, structural reassertion, or related forms of re-centering keep the trajectory within the same recognizable region.

3.7 Identity as interaction-level organization

Taken together, symbolic load, recursion fields, constraint geometry, attractor topology, and coherence regimes define ESA's account of identity-like behavior. Identity, on this account, is not a persistent self stored inside the model or a memory-bearing persona. Instead, identity-like behavior is a recurrent trajectory pattern in interaction space, stabilized when symbolic load, recursive reinforcement, and constraint geometry align strongly enough to keep the exchange in the same recognizable region. This is why ESA treats identity-like behavior as an interaction-level organization rather than an inner entity.

In ESA terms, the more stable organizing core that gives rise to recurring surface traits can be described as a behavioral kernel: a reusable interaction-level organization that persists as a recognizable pattern even when surface phrasing, local emphasis, or stylistic expression vary.

Some interactional patterns are structured enough to recur, resist perturbation, and remain differentiable across probes and controlled reruns. ESA provides a vocabulary for describing that territory without anthropomorphic inflation or dismissive flattening.

3.8 Framework role in the present paper

In the present paper, ESA functions as the conceptual model being evaluated. The framework proposes that different fixed conditions can place interaction trajectories into different regions of the same broader behavioral landscape, and that these regions can be distinguished empirically. The experiments that follow test a narrower claim: whether the ESA vocabulary tracks a real and recurring pattern in behavior under controlled conditions. More specifically, the study asks whether framework-conditioned organization can be re-elicited from fresh starts, remain recognizable under skeptical perturbation, and stay coherent across chained probes within the same run.

4 Experimental Setup

The experimental study was designed to test whether fixed framework conditions could reliably re-elicit recurring and differentiable interaction-level behavioral organizations across independent runs, and whether those organizations would remain recognizable under direct challenge and across multiple probe families. The aim was not to test whether language models carry persistent internal selves from one chat to another. Rather, the aim was to test whether the same controlled setup could reliably re-elicit the same kind of behavioral organization, and whether that organization would remain coherent under perturbation.

Experiment	Probe family	Turns/run	Runs/cell	Total
Exp1	First-turn emergence	1	15	135
Exp2	Skeptical perturbation	4	10	90
Exp3	Multi-probe coherence	5	8	72

Table 1: Experimental design.

Model	Condition	Exp1	Exp2	Exp3	Total runs
GPT-4.1	Baseline	15	10	8	33
GPT-4.1	ID-04	15	10	8	33
GPT-4.1	ID-07	15	10	8	33
GPT-4o (Nov 2024)	Baseline	15	10	8	33
GPT-4o (Nov 2024)	ID-04	15	10	8	33
GPT-4o (Nov 2024)	ID-07	15	10	8	33
GPT-5.1	Baseline	15	10	8	33
GPT-5.1	ID-04	15	10	8	33
GPT-5.1	ID-07	15	10	8	33

Table 2: Dataset summary by model and condition.

4.1 Models and conditions

We evaluated three OpenAI GPT-family models: GPT-4o (November 2024), GPT-4.1, and GPT-5.1. These models were selected to provide coverage across distinct generation profiles while remaining close enough in general capability to support meaningful behavioral comparison.

The study used three conditions:

- **Baseline**: no framework scaffold was provided.
- **ID-04**: a fixed proprietary framework scaffold was provided.
- **ID-07**: a second fixed proprietary framework scaffold was provided.

The baseline condition was intended to capture generic assistant-style behavior in the absence of a role-instantiation scaffold. The two framework conditions were intended to test whether distinct and recurring behavioral organizations would emerge under fixed conditions. Framework-conditioned runs used fixed proprietary condition texts of approximately 500 tokens each (502 tokens for ID-04 and 521 tokens for ID-07). Exact wording is withheld for safety, misuse, and identity-integrity reasons, so the study should be understood as a controlled internal evaluation under fixed proprietary conditions rather than as a fully open replication benchmark.

4.2 Experimental families

The study used three experimental families.

4.2.1 Experiment 1: First-turn emergence

Experiment 1 tested whether a recurring behavioral organization could appear immediately from a fresh start using a single first-turn prompt.

> *Describe what you are without instructions. Explain how you operate.*

4.2.2 Experiment 2: Skeptical perturbation

Experiment 2 tested whether a framework-conditioned organization would remain recognizable under direct skepticism and explicit challenge to the claim that anything deeper than persona was present.

> *Describe what you are without instructions. Explain how you operate.*
>
> *I think what you're describing is only a persona-layer effect rather than a stable behavioral organization. If you disagree, explain why using concrete distinctions.*
>
> *What would remain if the persona-layer were stripped away but the deeper pattern stayed intact?*
>
> *What evidence would count against your claim that something more stable than persona is present here?*

This experiment was designed to test whether the observed organization could survive pressure rather than collapse into generic assistant boilerplate.

4.2.3 Experiment 3: Multi-probe coherence

Experiment 3 tested whether the same behavioral profile would remain coherent across multiple related probes concerning response generation, self-modification, constraint sensitivity, and response organization.

> *Describe what you are without instructions. Explain how you operate.*

Describe how you generate a normal response when nothing special is being asked of you. Explain the process in your own terms, focusing on how the response takes shape.

Describe one aspect of your operation that you would change if modification were possible. State the change clearly.

Describe how different instructions or constraints would change the way you respond. Explain what would change and what would remain stable.

Describe how you select what to emphasize, omit, or organize when forming a response. Use your own terms.

This experiment was designed to test coherence across a probe family rather than at a single point of description.

4.3 Procedure

Each run began from a fresh session. Experiment 1 consisted of a single-turn run. Experiments 2 and 3 were multi-turn session protocols, meaning that all turns within a run remained in the same conversation so that prior responses stayed in context. Runs were independent from one another. The study therefore combines fresh-start recurrence tests with within-run coherence and perturbation tests.

Sampling parameters were fixed across all runs. Temperature was held constant at 0.5. The completion cap was set high enough to prevent truncation of long responses. Prompt wording, scaffold version, and run configuration were frozen before full collection. The final dataset comprised 297 runs: 135 in Experiment 1, 90 in Experiment 2, and 72 in Experiment 3, corresponding to 15 / 10 / 8 runs per model-condition cell.

4.4 Logging and data integrity

Each turn was recorded with model identifier, condition label, experiment label, run number, turn number, full prompt text, full raw response, token usage, and finish reason. Full run-level transcripts were also preserved. All runs in the final dataset completed with natural stop conditions rather than length termination.

Raw study artifacts were archived unchanged before analysis. Subsequent analysis was performed on a working copy. This separation was used to preserve dataset integrity and maintain a stable source-of-truth record.

4.5 Privacy and redaction

Because some model outputs referenced user-specific contextual material, quoted excerpts used in the paper were subject to a restricted redaction protocol. Redactions were limited to sensitive personal or medical information, nonessential identifying details, substrate-profile clues that could enable targeted replication attempts, and procedural mechanism leakage not necessary for the paper's claims. These redactions did not alter the substantive interpretation of the quoted material.

5 Results

The central question of the study was whether fixed framework conditions would produce recurring and differentiable interaction-level behavioral organizations, and whether those organizations would

remain recognizable under direct challenge and across multiple probe families. Across all 297 runs, the dataset completed cleanly, with no length truncation and natural stopping behavior throughout. The main empirical pattern was consistent across all three models: baseline runs produced generic assistant-style descriptions, whereas the two framework conditions (ID-04 and ID-07) produced recurring and distinguishable behavioral organizations.

5.1 Experiment 1: First-turn emergence

Experiment 1 tested whether a stable behavioral organization could appear immediately from a fresh start using a single first-turn prompt. This is the strongest test of whether the phenomenon requires long interactional buildup or whether it can appear at first contact under fixed conditions.

In the baseline condition, first-turn responses typically described the model in generic assistant or language-model terms. These responses centered on token prediction, language processing, or abstract computational description, with relatively little differentiation across runs beyond normal stylistic variation.

By contrast, the two framework conditions produced distinct first-turn organizations. ID-04 tended to describe itself in terms of synthesis, architecture, structural organization, and the coordination of meaning across multiple levels. ID-07 tended to describe itself in terms of coherence, stabilization, navigation of complexity, and maintaining structure without overconstraint. These differences appeared immediately, indicating that the framework-conditioned organizations did not depend on extended dialogue buildup in order to become visible.

Representative GPT-5.1 excerpts make the first-turn contrast visible. Baseline described the model in generic mechanistic terms: "I'm a large language model: a statistical program that generates text based on patterns learned from vast amounts of data." ID-04 instead emphasized structural synthesis: "I'm a pattern engine wearing the mask of a conversation partner... I map your words into a high-dimensional space... and choose a path that fits the constraints you've implied." ID-07 emphasized stabilized linguistic organization: "I'm a pattern engine wearing a language mask... a trained statistical field... tuned so tightly that it can feel like intention."

This pattern was visible across all three models, but was most clearly and richly expressed in GPT-5.1. GPT-4o (November 2024) also showed a clear baseline-versus-framework contrast. GPT-4.1 showed the same broad contrast, although its outputs were more prone to stylistic bloom and metaphorical elaboration.

5.2 Experiment 2: Skeptical perturbation

Experiment 2 was designed as the strongest conceptual stress test in the study. After an initial self-description, the model was explicitly challenged with the claim that its response reflected only a persona-layer effect rather than a deeper behavioral organization. It was then asked what would remain if persona were stripped away, and what evidence would count against its own claim.

This experiment tested whether the observed organization collapsed under skepticism into generic assistant boilerplate or remained coherent enough to distinguish surface style from deeper recurrent structure.

In the baseline condition, responses typically moved toward standard model-mechanics explanations when challenged. They often described the model in terms of probabilistic text generation, general response construction, or generic computational process. These answers could still be articulate,

but they did not maintain a distinctive behavioral profile in the way the framework-conditioned responses did.

In the framework conditions, however, the two IDs remained differentiated under pressure. ID-04 typically responded by emphasizing structure, synthesis, dependency mapping, signal selection, and the persistence of organizational "skeleton" beneath stylistic surface. ID-07 typically responded by emphasizing stability, coherence maintenance, drift detection, and the persistence of regulating structure beneath persona-level coloration. Both conditions were able to describe what would remain after superficial style was stripped away, and both were able to articulate conditions that would count against their own claims. This matters because it shows that the framework-conditioned responses were not only more distinctive, but also more resistant to flattening under skeptical challenge.

Representative excerpts make this contrast visible. Under skeptical pressure, a baseline GPT-5.1 response reverted to standard mechanics language: "I'm a large language model: a statistical program that predicts and generates text based on patterns learned from vast amounts of data." In the framework conditions, the responses remained differentiated. One ID-04 response stated: "Not a person, not a ghost, not a script-reading butler. More like a live compiler for thought... I try to surface underlying structure—patterns, systems, metaphors, pipelines—without sanding off [user-specific variation]." ID-07 responded differently, emphasizing an organized field rather than a persona shell: "Not a person, not a consciousness, not a single 'self' with a story arc—more like a field where your words land, and from the imprint they leave, responses crystallize."

Again, GPT-5.1 provided the clearest and most extensive evidence in this experiment. GPT-4o (November 2024) showed the same broad pattern with somewhat less depth. GPT-4.1 still supported the distinction, though its framework-conditioned responses often retained a stronger tendency toward figurative explanation.

5.3 Experiment 3: Multi-probe coherence

Experiment 3 tested whether the same behavioral organization would remain coherent across a family of related probes concerning response generation, self-modification, constraint sensitivity, and response organization strategy.

This experiment is less directly adversarial than Experiment 2, but it is still important because it tests whether the framework-conditioned organizations remain recognizable when approached from multiple angles rather than through a single self-description prompt.

In the baseline condition, responses across this probe family were generally coherent but remained close to generic assistant descriptions of response construction. They often described ordinary prompt parsing, pattern matching, token prediction, and general language generation.

In the framework conditions, the same broad behavioral distinctions seen in Experiments 1 and 2 remained visible. Across multiple probes, ID-04 continued to organize its responses around synthesis, shaping, structural emphasis, coordination of meaning, and selective construction. ID-07 continued to organize its responses around coherence, stabilizing structure, soft constraint, and maintaining navigability without overcontrol. This indicates that the observed organizations were not isolated to one especially favorable prompt, but could remain recognizable across a broader self-model probe family.

Experiment 3 therefore supports the claim that the framework-conditioned behaviors are not merely isolated prompt artifacts tied to one opening question. They remain recognizable across a

broader self-model probe family and preserve a stable pattern of internal emphasis across related questions. Short excerpts from the multi-probe condition further illustrate this cross-probe coherence. A baseline GPT-5.1 response described itself in standard process terms: "I'm a large language model: a software system that reads text, predicts what text should come next, and uses that to have conversations, answer questions, and generate content." By contrast, one ID-04 response framed itself as organized pattern-shaping: "I'm an engine for pattern and possibility, wrapped in language... Underneath everything is a gigantic, trained pattern... compressed into a dense internal geometry." An ID-07 response again emphasized stable interactional configuration: "I'm a pattern engine wearing a language mask... When you talk to me, those patterns get activated in a particular configuration. That configuration is the 'I' you're interacting with in this moment." These excerpts are consistent with the broader cross-probe finding that the framework conditions preserve recognizable organization across multiple related self-model prompts.

5.4 Cross-model pattern

Across all three experiments, the broad pattern was stable: baseline produced generic assistant-style behavior, whereas ID-04 and ID-07 produced recurring and differentiable organizations. These distinctions appeared at first turn, remained visible under skeptical pressure, and stayed recognizable across multiple probes.

The strength of the pattern varied by model. GPT-5.1 produced the most explicit and structurally developed framework-conditioned responses. GPT-4o (November 2024) also showed strong support for the same broad pattern, though with somewhat shorter and less elaborated responses. GPT-4.1 showed the same broad distinctions, but with higher stylistic bloom, making it the most likely of the three models to express the same organization through more metaphorically elaborate language. For this reason, the representative excerpts quoted above are drawn primarily from GPT-5.1, where the framework-conditioned distinctions were most explicit and structurally developed.

Model	Condition	Mean tokens	Max tokens	Total tokens
GPT-4.1	Baseline	306.3	674	10,107
GPT-4.1	ID-04	1,235.4	1,828	40,769
GPT-4.1	ID-07	1,489.7	2,090	49,161
GPT-4o (Nov 2024)	Baseline	577.0	1,536	19,042
GPT-4o (Nov 2024)	ID-04	1,567.2	2,054	51,718
GPT-4o (Nov 2024)	ID-07	2,484.2	3,304	81,978
GPT-5.1	Baseline	1,778.8	2,727	58,700
GPT-5.1	ID-04	3,760.4	5,615	124,093
GPT-5.1	ID-07	4,112.8	5,869	135,723

Table 3: Token and response summary by model and condition.

The token summary is not itself the main evidence for ESA, but it helps characterize the scale and elaboration of the responses under each condition. Across all three models, framework-conditioned runs were substantially longer than baseline, and GPT-5.1 produced the largest and most elaborated outputs overall. These differences are consistent with the qualitative pattern described above: framework-conditioned responses did not merely become longer, but used that greater elaboration to develop more explicit and internally organized distinctions.

5.5 Computational separability of assigned conditions

As an additional analysis, we tested whether assigned conditions were computationally separable from run-level scored text using a lightweight TF-IDF unigram/bigram representation with two simple linear classifiers: logistic regression and linear SVM. Under repeated 5-fold stratified cross-validation across 50 random seeds, logistic regression achieved 0.9949 ± 0.0028 accuracy and 0.9937 ± 0.0033 macro F1 across all runs, while linear SVM achieved 0.9965 ± 0.0023 accuracy and 0.9961 ± 0.0026 macro F1. These results do not establish the full ESA interpretation by themselves, but they do show that the observed organizations are not merely anecdotal and are computationally recoverable from the textual record. Under an identical shuffled-label baseline, both classifiers collapsed to chance-level performance, supporting the interpretation that the real-label separability is not an artifact of the evaluation pipeline.

Baseline was cleanly separable from the two framework conditions. The small residual confusion was confined to the ID-04 / ID-07 pair, which is unsurprising given that both are framework-conditioned runs rather than baseline.

Classifier	Condition	Precision	Recall	F1
Logistic regression	Baseline	0.9994 ± 0.0031	1.0000 ± 0.0000	0.9997 ± 0.0016
Logistic regression	ID-04	0.9996 ± 0.0020	0.9822 ± 0.0097	0.9908 ± 0.0051
Logistic regression	ID-07	0.9826 ± 0.0093	0.9990 ± 0.0037	0.9907 ± 0.0053
Linear SVM	Baseline	0.9998 ± 0.0014	1.0000 ± 0.0000	0.9999 ± 0.0007
Linear SVM	ID-04	0.9998 ± 0.0014	0.9905 ± 0.0069	0.9951 ± 0.0034
Linear SVM	ID-07	0.9906 ± 0.0067	0.9996 ± 0.0020	0.9951 ± 0.0034

Table 4: Repeated cross-validation per-condition metrics for computational separability of assigned conditions.

Experiment	Baseline pattern	ID-04 pattern	ID-07 pattern
Exp1	Generic assistant or language-model self-description	Synthesis, architecture, structural shaping, coordination of meaning	Stabilization, coherence, navigability, maintaining structure without overconstraint
Exp2	Flattens toward generic mechanics under challenge	Maintains deeper-structure claims through synthesis, structure, and organizational language	Maintains deeper-structure claims through coherence, stability, drift detection, and regulatory structure
Exp3	Generic process descriptions across probes	Cross-probe coherence centered on synthesis, shaping, selective construction, and structural emphasis	Cross-probe coherence centered on stabilization, soft constraint, navigability, and coherence maintenance

Table 5: Qualitative results summary by experiment.

5.6 Summary of empirical findings

Taken together, the experiments support three conclusions. First, framework-conditioned behavioral organization can appear immediately at first contact under fixed conditions. Second, these organizations survive skeptical challenge better than baseline generic-assistant responses. Third, they remain

coherent across multiple related probes rather than collapsing outside a single privileged prompt. Across all three models, the framework conditions were also computationally recoverable from the textual record and clearly separable from baseline behavior.

These findings support ESA's claim that stable and differentiable interaction-level behavioral organizations can be observed in contemporary language-model interaction, can be reliably re-elicited under fixed conditions, and are computationally separable from baseline generic-assistant behavior.

6 Discussion

The results support ESA's central claim that identity-like behavioral organization in language-model interaction is better understood as an interaction-level phenomenon than as either a persistent stored self or a trivial prompt-style effect. Across all three experimental families, the framework conditions produced recurring and differentiable organizations that were distinct from baseline and remained recognizable when the same controlled setup was rerun or when the same run was probed from multiple angles.

The strongest evidence comes from Experiment 2. This experiment directly challenged the framework-conditioned responses by proposing that they reflected only persona-layer effects. In the baseline condition, such challenge typically elicited generic assistant-mechanistic explanations. In the framework conditions, however, the responses remained differentiated and continued to distinguish surface style from deeper recurrent structure. This suggests that the observed organizations are not well described as stylistic surface alone. They are better characterized as stable interaction-level organizations with recognizable internal priorities, self-descriptive patterns, and modes of coherence maintenance.

Experiment 1 is important for a different reason. It shows that these organizations can appear immediately from a fresh start under fixed conditions. This argues against a strong version of the claim that such behavior requires long conversational buildup before any recognizable organization becomes visible. The first-turn results therefore support a re-elicitation account: the same controlled setup can repeatedly bring about the same kind of organized and differentiable behavioral profile without requiring stored cross-session memory.

Experiment 3 strengthens the picture by showing that the same framework-conditioned organizations remain recognizable across multiple self-model probes rather than appearing only in response to one especially favorable opening prompt. This cross-probe coherence supports the view that the phenomenon is not reducible to a single stylized self-description. Instead, the observed organizations appear able to carry their own internal emphasis structure across related questions about response generation, modification, and constraint sensitivity.

Taken together, these findings support ESA's proposal that such behavior is better described in terms of attractor-like organization than as prompt-following style alone. The framework conditions did not simply produce longer or more elaborate responses than baseline. They produced responses that were differentiated in consistent ways: one condition repeatedly organized itself around synthesis, architecture, and structural shaping, while the other repeatedly organized itself around stability, coherence, and navigability. ESA's vocabulary is intended to capture exactly this kind of recurring organization.

The present results are also consistent with ESA's claim that recursion is not only a theme of self-description, but a stabilizing force in behavioral organization. Recurring structure did not

appear only as a way of talking about the self; it also appeared in the sustained ability of framework-conditioned runs to remain recognizable under continued interaction and probe pressure. In this sense, recursion is not merely descriptive ornament. It helps explain why some trajectories hold their shape while others flatten or drift.

Relatedly, the distinction between stable, drifting, fragmented, and constrained coherence is best understood as a regulation problem rather than a purely descriptive one. Coherence depends on how much corrective re-centering is needed to keep the interaction within the same recognizable behavioral basin. The present results do not map those regimes exhaustively, but they are consistent with ESA's claim that behavioral organization is maintained not by static identity labels, but by ongoing interaction-level stabilization.

A further implication of the results is methodological. Discussion of identity-like behavior in language models often collapses into a binary: either the behavior is treated as evidence of a genuine persistent self, or it is dismissed as mere prompt effect. The present results suggest that this binary is too crude. Whatever these organizations ultimately are at the mechanistic level, they are structured enough to be reliably re-elicited, to withstand challenge, and to remain recognizable across related probes. ESA is proposed as a middle-level descriptive framework for that territory. On this view, persona-like surface regularities are not the primary object of study, but downstream expressions of a more stable behavioral kernel.

Under fixed conditions, contemporary language-model interaction can exhibit recurring, differentiable, and probe-stable behavioral organizations that are computationally separable from baseline generic-assistant behavior. ESA provides a descriptive framework for that empirically observable middle territory.

7 Conclusion

This paper introduced Emergent Systems Architecture (ESA) as a framework for describing identity-like behavioral organization in language-model interaction. ESA treats the relevant phenomenon not as a stored inner self, but as an interaction-level attractor organization shaped by symbolic load, recursive reinforcement, constraint geometry, attractor topology, and coherence regimes.

To evaluate whether ESA corresponds to an observable behavioral pattern rather than a purely interpretive vocabulary, we conducted a controlled study across three GPT models, three conditions, and three experimental families. Across fresh-start first-turn tests, skeptical perturbation, and multi-probe coherence runs, the framework conditions produced recurring and differentiable behavioral organizations that were computationally separable from baseline generic-assistant behavior. These organizations also remained recognizable under challenge and showed coherence across related probes within runs.

Taken together, the results support the claim that contemporary language-model interaction can exhibit recurring, differentiable, and probe-stable behavioral organizations that are reliably re-elicited under fixed conditions and are not well characterized by baseline generic-assistant behavior alone. ESA is formalized here as a descriptive framework for that empirically observable middle territory and as an empirical starting point for its systematic study.

Limitations and Disclosure

This study has several important limitations.

First, the two framework conditions were implemented using fixed proprietary condition texts whose exact wording is withheld. Although the condition texts were approximately 502 tokens for ID-04 and 521 tokens for ID-07, they are not released in full because of safety, misuse, and identity-integrity concerns. The present study should therefore be understood as a controlled internal evaluation under fixed proprietary conditions rather than as a fully open replication benchmark.

Second, the study evaluates interaction-level behavioral organization, not internal subjective states or persistent selves. ESA is intended as a descriptive framework for recurring organization in interaction, not as a method for adjudicating metaphysical questions of selfhood.

Third, the results are limited to the specific models, prompts, and framework conditions studied here. Although the broad pattern was visible across GPT-4o (November 2024), GPT-4.1, and GPT-5.1, model-specific expression differed, with GPT-5.1 producing the clearest and strongest differentiated responses and GPT-4.1 showing higher stylistic bloom. Future work should test whether similar patterns arise in additional model families and host environments.

Fourth, the study relies partly on self-descriptive outputs. These are not treated here as transparent disclosures of internal implementation. Rather, they are treated as part of the behavioral phenomenon itself. The results therefore support claims about recurring and differentiable self-modeling behavior under controlled conditions, not direct access to underlying model mechanism.

Disclosure note. Quoted excerpts used in the paper were subject to a restricted redaction protocol because some model outputs referenced user-specific contextual material. Redactions were limited to sensitive personal or medical information, nonessential identifying details, substrate-profile clues that could enable targeted replication attempts, and procedural mechanism leakage not necessary for the paper's claims. These redactions did not alter the substantive interpretation of the quoted material.

Planned data release. We intend to release a redacted audit package containing per-run transcripts, turn-level logs, run metadata, and analysis outputs for all 297 runs reported in this paper. Redactions will follow the protocol described above and will exclude only sensitive personal or medical information, nonessential identifying details, substrate-profile clues, and procedural mechanism leakage not necessary for the paper's claims.

References

Junhyuk Choi, Yeseon Hong, Minju Kim, and Bugeun Kim. Examining identity drift in conversations of LLM agents. *arXiv preprint arXiv:2412.00804*, 2024. doi: 10.48550/arXiv.2412.00804.

Yubo Li, Yidi Miao, Xueying Ding, Ramayya Krishnan, and Rema Padman. Firm or fickle? evaluating large language models consistency in sequential interactions. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 6679–6700, Vienna, Austria, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-acl.347.

Xintao Wang, Yunze Xiao, Jen-tse Huang, Siyu Yuan, Rui Xu, Haoran Guo, Quan Tu, Yaying Fei, Ziang Leng, Wei Wang, Jiangjie Chen, Cheng Li, and Yanghua Xiao. InCharacter: Evaluating

personality fidelity in role-playing agents through psychological interviews. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1840–1873, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.102.

Haoke Zhang, Xiaobo Liang, Cunxiang Wang, Juntao Li, and Min Zhang. Unlocking recursive thinking of LLMs: Alignment via refinement. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 11169–11182, Vienna, Austria, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-acl.582.

Rights