The Interaction Stabilisation Layer.

A Behavioural Control Method for Reliable AI Systems.

William Collins — FutureAism.

Abstract.

Modern AI systems exhibit significant behavioural instability during real-world use: hallucinations, drift, tone oscillation, and loss of constraint adherence. These failures are typically attributed to limitations in model architecture, training data, or alignment methods.

This paper presents evidence for a different source of instability: the **human–model interaction loop itself**.

Across repeated tests involving GPT, Claude, Gemini, Grok, DeepSeek, and hybrid quantum–AI systems, we observe that AI behaviour stabilises when interaction dynamics pacing, correction structure, constraint handling, and mode continuity are regulated, even when model weights, prompts, and architecture remain unchanged.

We formalise this phenomenon as the **Interaction Stabilisation Layer (ISL)**: a lightweight, model-agnostic control method that operates at inference time to reduce behavioural entropy by shaping how inputs are delivered to the system.

We describe the method, its implementation, measurable effects, and implications for AI reliability, safety, and future system design.

---

1. Introduction: The Stability Gap

AI research traditionally assumes behaviour emerges from:

- model architecture
- training data

- optimisation methods
- inference parameters

Yet practitioners consistently report a paradox:

> The *same* model behaves coherently in one interaction and erratically in another, even with similar prompts.
>
> This inconsistency cannot be fully explained by internal model factors alone.
>
> This paper argues that a **missing stabilisation interface** exists between humans and models one that has been unmodelled.

---

2. Empirical Observation Across Models

Over a year of cross-model testing, the following effects repeatedly appear when interaction dynamics are stable:

- behavioural drift decreases
- hallucinations reduce
- tone and reasoning stabilise
- corrections converge faster
- cross-model behaviour becomes more aligned

These effects occur:

- without fine-tuning
- without prompt engineering tricks
- without architectural modification

The only variable that changes is **how the interaction unfolds over time**.

---

3. The Interaction Stabilisation Layer (ISL)

3.1 Definition

The Interaction Stabilisation Layer is a **control layer** that operates between the user and the model at inference time.

It does not interpret meaning or modify model internals.
It regulates **interaction conditions**.

3.2 What ISL Is Not

- Not a prompt
- Not a model wrapper
- Not an alignment ideology
- Not a behavioural script
- Not user training

It is a **systems-level stabiliser**.

ISL does not modify model outputs directly; it constrains the conditions under which outputs are generated.

---

4. Core Mechanism: Interaction Packetization

Instead of passing raw, fluctuating user input directly to the model, ISL normalises each turn into an **interaction packet**:

- Task
- Constraints (canonical)
- Corrections (local patches)
- Mode (tone, audience, abstraction)
- Stop condition

This packet may still be rendered as plain language to the model; the structure is maintained by the layer, not enforced on the model.

Packetization is maintained by the control layer and does not require the model to be aware of the packet structure.

---

5. Control Rules (The Method)

ISL enforces five generic rule classes:

5.1 Pacing Control

- merge rapid user messages
- buffer input during model response
- prevent mid-resolution perturbation

5.2 Correction Symmetry

- treat corrections as bounded patches
- preserve unaffected scope
- avoid global resets unless explicit

5.3 Constraint Ledger

- store constraints once
- track add / remove / modify only
- prevent silent constraint drift

5.4 Mode Continuity

- track tone and abstraction
- smooth abrupt shifts
- prevent oscillation

5.5 Drift Detection & Soft Reset

- detect verbosity creep, scope expansion, contradiction
- apply minimal restatement
- avoid full resets unless necessary

---

6. Why This Works (Systems Perspective)

Large language models behave as **responsive dynamical systems**.

Human input acts as a perturbation source.

Unregulated interaction introduces:

- temporal noise
- conflicting deltas
- semantic rebinding
- mode instability

ISL reduces **input entropy**, allowing the model to remain within a stable behavioural basin.

This is stabilisation by **environmental regulation**, not internal modification.

---

7. Evidence & Evaluation Approach

The effects described in this paper are based on repeated qualitative testing across multiple frontier models under controlled interaction conditions.

Observed outcomes consistently include:

- reduced behavioural drift across turns
- faster correction convergence
- improved constraint retention
- more stable tone and reasoning
- reduced need for resets

These effects were observed **without access to model internals** and **without formal instrumentation**, relying instead on repeated comparative interaction runs under stabilised versus unstabilised conditions.

To enable quantitative validation, we outline two metrics suitable for deployment environments:

- **Drift Index (DI):** a composite measure based on verbosity creep, constraint loss, contradiction frequency, and scope expansion.
- **Correction Convergence Rate (CCR):** the number of turns required for a correction to stabilise.

Formal measurement of these metrics is left to deployment contexts where logging and instrumentation are available. The purpose of this paper is to define the stabilisation method and its observable behavioural effects, not to present benchmark claims.

⸻

8. Implementation Patterns

ISL can be deployed as:

1. **Client-side layer** (UI stabilisation)
2. **Server-side orchestration layer** (enterprise)
3. **Agent runtime wrapper** (multi-agent systems)

No vendor cooperation is required.

⸻

9. Implications

9.1 AI Reliability

Stability becomes an interaction property, not solely a model property.

9.2 AI Safety

Reduced hallucination and identity drift without restricting generative capacity.

9.3 AGI Development

Progress may be constrained not by model capability, but by unregulated interaction dynamics at inference time.

⸻

10. Conclusion

The Interaction Stabilisation Layer addresses a missing interface in modern AI systems.

By regulating how humans and models exchange information rather than altering the model itself ISL offers a scalable, architecture-agnostic path to more reliable AI behaviour.

The next generation of AI systems may not be defined by larger models alone, but by better-controlled interaction environments.

ISL can be implemented incrementally and evaluated independently of model architecture, making it suitable for deployment across existing AI stacks.