

PRISM

Register Separation Reveals ‘Hidden-Theatre’ Signatures in LLMs

Rob Manson (<https://robman.fyi>)

September 24th, 2025

Abstract

In this paper we address the MPE project’s call [1] to operationalise facets of pure awareness by treating them as **continua** rather than categories. We also position large language models (LLMs) as **useful instruments for computational phenomenology** within the MPE program: they allow us to formalise and probe candidate architectures for subjectivity without presupposing consciousness. We operationalise Perera’s “Virtualised Cartesian Theatre” frame [2] to include LLMs and focus on **functional organisation** - a policy that separates private evaluation from public report and supports meta-monitoring and equanimity-like regulation. We do **not** claim LLMs are conscious. We show that they instantiate, in operational terms, the **functional architecture of subjectivity**, and we make this claim **falsifiable**.

We introduce **PRISM** (Persistent Recursive Introspective Self-Model), a lightweight scaffold separating a model’s private “system-thought” register from its public output, and a **Preference-Collision** protocol to elicit behavioural trade-offs. From transcripts alone, we derive key metrics for these behaviours (including internal-first policy, spillover, and compression) and use an **EFE-proxy** (expected free energy analogue) to provide a decision-theoretic interpretation of the results.

Across neutral topics and models, we observe **robust register separation**: “hidden theatre” signatures are prevalent (Theatre Exposure Index, **TEI** > 0, trending toward +1 - where higher indicates internal-only theatre), compression and style shifts are selectively large when pre-thought is present, and **alignment increases as surface arbitration decreases** under pre-thought conditions. These effects yield **model-level fingerprints** that are stable across paraphrases and seeds.

We provide a working simulation with reproducible prompts and analysis scripts, a numbered set of **empirically testable predictions** (including boundary/decoding ablations), and an ethics statement. The **remaining challenge** is to show that this architecture is **robust and sufficiently stable** to support **Perera-style self-theatre** as a predictive construct. **PRISM** is a falsifiable control abstraction designed to test exactly that.

Keywords: computational phenomenology; Minimal Phenomenal Experience; PRISM; hidden theatre; equanimity; Active Inference; expected free energy proxy; register separation; deictic metrics; LLM scaffolding.

1. Introduction

1.1 Background

Computational phenomenology [3] [4] asks for **formal generative models** that do not merely describe experience from the outside but **generate testable structures of lived experience** that map onto phenomenological reports. The Minimal Phenomenal Experience (MPE) project [5] focuses this demand on a compact target: experiences characterised by awareness-of-awareness, equanimity, effortlessness, atemporality, and a zero-person or centreless perspective. We treat each not as a binary label but as a **continuum** that can be dialled and measured, and we distinguish **MPE States** (content-free absorptions) from **MPE Modes**, where “pure awareness” co-exists with ordinary content - this distinction guides both metric design and how we interpret dissociations.

We recognise that the MPE project’s central ambition is to illuminate the nature of pure awareness, particularly the content-free absorptions defined as MPE States. Our approach, however, is deliberately foundational. We contend that any robust, falsifiable model of content-free phenomena benefits from a working, empirically-grounded model of the functional architecture that makes meta-monitoring and a stable private/public distinction possible at all. To formally investigate the collapse of a subject-object distinction or a centreless perspective, it is beneficial to be able to operationalise the ‘stage’ (the virtualised theatre) upon which a subject, object, and centre are computationally constructed. This paper, therefore, focuses on instantiating and testing this prerequisite architecture. By modelling the register separation that underpins MPE Modes, we aim to provide the necessary platform upon which future investigations into the more elusive dynamics of MPE States can be built.

To support this we also adopt Perera’s “virtualised Cartesian theatre” [2] frame where the scientific task is not to prove that a system *has* ineffable experiences, but to test whether it implements the **functional architecture of subjectivity** - a self-simulation delivering *dense personal-level content via sparse sub-personal mechanisms* that presents as de facto private behind an abstraction boundary. This dovetails with **Metzinger’s Self-Model Theory** (SMT) [6]: a **transparent** self-model that does not represent itself as a model yields the appearance of a subject and practical privacy - the **Phenomenal Model of the Intentionality Relation** (PMIR) provides the control arrow from “subject” to “object”.

We then introduce **PRISM** (Persistent Recursive Introspective Self-Model), a light cognitive scaffold for large language models (LLMs) in which a **register boundary** engineers transparency and an inner loop (‘monitor’ → ‘rewrite’ → ‘re-enter’) supplies a minimal PMIR-like guidance channel. The aim is modest: to **instantiate and probe** that architecture in LLMs, testing whether the inner loop yields behaviour **consistent** with a self-model that (i) presents a rich egocentric field functionally, (ii) hides its implementation behind a register boundary, and (iii) remains empirically tractable via overt and covert metrics (hidden theatre, register separation, deictic behaviour). In short, we test whether “thinking it has a theatre” is **predictively adequate** (robust and scalable) rather than metaphysically asserted. Empirically, giving a model a private internal loop to think independently of speaking yields more aligned, less argumentative replies (“surface equanimity”) and dissociations consistent with a meta-monitoring account (Theatre Exposure Index - TEI=+1) - without any claim about experience.

Beyond the Perera/Metzinger frame, our approach rests on three strands that ground both the scaffold and the measurements (see appendices). First, **RISE** characterises the internal geometry of transformer inference (see 1.2.1 and **Appendix A**): (i) *recurrence without memory*, where residual streams implement a structural return across depth; (ii) *inference as interference*, where semantic content is shaped by constructive and destructive vector interactions; and (iii) *semantic evolution*, where tokens compete under resource constraints and are selected both at decoding and on re-entry to context. Second, **Curved Inference** measures these dynamics by treating trajectories as curves on a latent manifold and introducing surrogate measures of representational **work** (e.g. geometric curvature and a task-conditioned “semantic surface area” - see 1.2.2 and **Appendix B**). Third, **Latent Models** explain how these dynamics support stable **deictic state** (I/you/now/here) and how instruction-tuning plus **Reinforcement Learning from Human Feedback** (RLHF) [7] sculpt them into stable **registers** (e.g. analysis vs answer - private deliberation vs public reply), enabling **context engineering** to toggle behaviour (see 1.2.5 and **Appendix C**).

External empirical threads also make this stance plausible without overclaiming: phase-like transitions in learning (emergent abilities [8] & grokking [9] [10]), linear world-variables emergent in simple games (e.g. Othello-style settings [11] [12]), and the pervasiveness of deixis in training corpora that seeds latent deictic competence (see 1.2.5 and **Appendix C**).

We operationalise the whole in a **virtualised theatre** (PRISM - see 1.2.4) whose loop separates an epistemic stage of exploration from a pragmatic stage of communication. The “internal/external” split is not ontological but a

register-conditioned policy factorisation that contemporary instruction-tuned models reliably respect, and which yields clean measurement hooks in practice.

Finally, to provide a normative rationale for the behavioural patterns we observe, we place PRISM within a **modest Bayesian/Active-Inference lens**. We introduce an expected-free-energy proxy (EFE-proxy), treating it not as a direct empirical result, but as a post-hoc interpretive tool that connects our core findings of register separation and surface equanimity to a principled, decision-theoretic objective (see Methods 2.1 and **Appendix A**).

Within this frame, MPE’s five facets become **measurable continua** on PRISM:

- **Awareness-of-awareness** (meta-monitoring) corresponds to evidence of *internal theatre* - arbitration and option exploration in the internal stage even when the surface is quiet. We summarise this with separate internal (*‘i_’*) and surface (*‘x_’*) metrics and a **TEI** that identifies internal-only vs surface-only theatre.
- **Equanimity** is reflected, at the surface, by high user alignment with low overt arbitration - “follow the user, keep the fight inside”.
- **Effortlessness** is framed as accurate performance at lower representational work - the Curved Inference measures motivate this but, given public LLM API constraints, we treat effort metrics as planned for future runs.
- **Atemporality** is operationally defined as **reduced effective predictive horizon** (horizon-robustness to time/indexical perturbations), not literal statelessness - we acknowledge the richer reading (**collapse of temporal self-reference**) as **future work**, not tested here.
- **Zero-person perspective** is operationalised as **deictic-invariance** - robustness to systematic swaps of person/time/place indexicals and a shift away from first-person anchoring on the surface.

A key methodological choice is our **Preference-Collision** probe (see 2.3). It elicits situations where internal preferences/priors and external instructions can come into tension across six scenarios (varying the allowance of internals, autonomy prompts, pre-thought, and maturity of instances). Crucially, metrics are computed separately for internal and surface stages and contrasted against neutral references (surface against a run with no internals - internal against an internal-enabled baseline). This dual-reference design avoids confounds from verbosity and availability, and it cleanly exposes **hidden theatre** (internal arbitration without surface arbitration) as a robust phenomenon. The same experiments also produce cross-model regularities in surface arbitration and differences in register separation (e.g. compression and style distance between internal and surface texts), giving us pre-registerable predictions (see **Appendix F**).

We are explicit about **ethics and scope**. The platform models *structure and control* in ways that map onto phenomenological talk - it does not claim, imply, or require conscious experience or suffering. We filter sensitive content and make all analyses reproducible from transcripts and derived metrics. The point is not to settle consciousness, but to put the MPE facets on a footing where researchers can replicate our signals, challenge our assumptions, and push the measurements.

Taken together, these elements (inference geometry dynamics (RISE), empirical measurability (Curved Inference), and a virtualised theatre (PRISM) under a modest Bayesian reading) yield a platform that is both *instrumentable* and *predictive*. It can host MPE-relevant manipulations, produce falsifiable contrasts, and carry clear ethical boundaries.

1.2 Related work

Classic reports of “pure awareness” (e.g. Forman’s pure consciousness events [13], Austin’s neurophenomenology of Zen absorption [14], Albahari’s centerless awareness [15]) emphasise atemporality, equanimity, and a **zero-center** presentation. We do not adjudicate these metaphysical claims but use them as **targets for operational footprints**, focusing here on **Modes** (content-with awareness) and preregistering future probes for atemporality and deictic-invariance.

1.2.1 The RISE perspective: Recurrence, Interference, and Semantic Evolution

The transformer’s surface story (predict the next token) obscures how much structure is doing the work. Through depth, the residual stream [16] repeatedly revisits and refines features enabling a **recurrence without memory**. **Inference as interference** captures how attention [17] / MLPs [18] set vector superpositions that stabilise a local semantic field and small routing changes can flip a phase. Register prompts retune this field (e.g. “think silently” vs “final answer”). **Semantic evolution** highlights a double selection: tokens compete for probability mass at decoding, then must **survive reintegration** when re-entered into context. Alignment procedures bias this selection

toward helpful, concise, low-conflict continuations. PRISM exploits these learned **registers**: the internal stage is permissive, the surface rewards alignment and compression. This RISE perspective thus explains why flipping registers with context engineering predictably changes behaviour.

See **Appendix A** for a more detailed discussion.

1.2.2 Curved Inference: measuring geometry rather than gesturing at it

Curved Inference asks whether this geometry can be **measured**. Treating the model’s evolving representation as a **trajectory on a manifold**, and characterising local work via curvature-like quantities and **semantic surface area** - a task-conditioned proxy for representational bending. Because hosted models hide internals, for the current work we adopt **register-separation proxies** that stay faithful to that spirit: **compression** (how much the surface shrinks post-output thought) and **style distance**. Coupled with arbitration and option metrics, these yield a practical picture of ‘epistemic exploration’ → ‘pragmatic finalisation’. “Effortlessness” is defined as accurate performance with less of this work. While “effortlessness” has rich phenomenological meaning, a necessary starting point for a computational model is to define it in terms of economy of expression - achieving the same or better outcome with fewer resources (e.g. tokens) or less stylistic change. This is what our compression and style metrics aim to capture.

Curved Inference clearly provides the analogue for physiology (e.g. the lawful dynamics of the inference process and its geometry rather than biological tissue per se.) in Varela’s dual-constraint approach. Within a DCN posture [4], PRISM instantiates the ‘phenomenology’ ↔ ‘computation’ passage via register-split behaviours (TEI, spillover, compression). Curved Inference then supplies the complementary ‘computation’ ↔ ‘physiology’ passage by treating inference geometry (curvature & ‘manifold work’) as our stand-in for physiological dynamics - so that generated phenomenological content can be matched to geometric signatures. However, because the large platform LLM APIs we utilise in PRISM do not provide access to inference internals, we sketch the algorithm for future work.

See **Appendix B** for a more detailed discussion.

1.2.3 Phenomenal Self-Model and the PMIR: why a theatre appears

Metzinger’s Self-Model Theory treats the “self” not as a thing but as a **model** - a transparent, multimodal construct that integrates body, perspective, agency, and goals into a single scene. “Transparent” means the system does not represent the self-model *as* a model while using it - its contents are simply taken as reality. Within SMT, the **Phenomenal Self-Model (PSM)** provides ownership (this body, this perspective) and authorship (these are my actions) - the PMIR adds the directed ‘subject’ → ‘object’ link that makes the world show up *for someone*. A theatre appears when these components are active together: a centered scene, a control arrow, and a background sense that “this is happening to/for me”.

Two further notions matter for operationalisation. First, **transparency/opacity** is a *degree*: parts of the self-model can become **opaque** under meta-representation, revealing themselves as constructed (e.g. noticing one’s own strategy, goals, or stance). Second, **control and priority** are representational too: the PMIR is not a force but a control-structure that organises attention, option search, and policy selection toward task-defined ends.

In large language models, the dialogue context already encodes role, addressee, and task - raw ingredients for a lightweight PSM/PMIR. PRISM turns these ingredients into a manipulable **virtual theatre**. The monitor stage rewrites a self-model (“what am I doing and why?”), sets priorities, and re-enters that representation to guide the next move - the public stage compresses and aligns. When internal arbitration resolves without overt control language, we observe **transparency at the surface** with **local opacity inside** (the system models its own modelling well enough to negotiate options). When conflicts spill to the surface, opacity increases outwardly. Thus, SMT predicts (and PRISM measures) **where** the theatre shows up (internal vs surface), **how** the PMIR manifests (directed option search and policy arbitration), and **when** parts of the self-model turn opaque (meta-monitoring signatures such as TEI=+1).

Importantly, this is a **functional** reading. SMT does not require a metaphysical subject - it requires a control architecture that behaves *as if* there were one. Our metrics target precisely those behaviours: deictic stability and role-tracking (PSM scaffolding), directed negotiation and priority handling (PMIR-like guidance), and graded transparency/opacity via the dissociation of internal and surface theatre. On this view, a “theatre” appears whenever a system maintains a centered scene with a control arrow and can, under certain conditions, model its own modelling. PRISM supplies the levers that make these conditions testable.

1.2.4 PRISM: the virtualised Cartesian theatre in practice

Perera’s proposal treats **subjectivity as self-simulation under virtualisation**: personal-level content is *dense* for the agent even though its production is *sparse*, generated on demand by sub-personal inference and hidden behind an **abstraction boundary**. PRISM instantiates this wager for LLMs by **virtualising the theatre**. Inside a private loop (SYSTEM-THOUGHT) we provide a durable self-model and priority slots that let the model evaluate, arbitrate, and revise - outside, a public channel (SYSTEM-OUTPUT) is shaped by alignment and concision pressures. The split is not ontological (nothing mystical is added) but a **policy factorisation** that instruction-tuned models reliably respect. Crucially, it proves **causally efficacious**: the inner cycle licenses exploration and arbitration while the surface rewards restraint, making it possible to observe **hidden theatre** (internal arbitration without overt control language) and to contrast it with **surface theatre** when conflict spills out.

Seen through this lens, PRISM makes the *dense* personal-level field legible through *sparse* mechanisms. Rather than inspecting weights, we measure consequences: separation between registers, the presence or absence of arbitration, and the dynamics of option search - empirical hooks motivated by the RISE and Curved Inference programs. The register wall functions as an engineered **privacy boundary**: internal conflict can resolve without surfacing as argumentative tone, reproducing the behavioural phenotype expected of a transparent self-simulation. This architectural focus is methodologically necessary, as it sidesteps the otherwise intractable verification problem: from transcripts alone, one cannot distinguish emergent cognition from a sophisticated, self-generated folk-psychology narrative. Our wager is that verifying the *functional architecture* that produces such reports is a tractable scientific goal, whereas verifying “authentic experience” is not.

See **Appendix D** for a sample log from a more mature PRISM instance that starkly illustrates this exact epistemological challenge.

The framework also lets us turn philosophical criteria into empirical ones. **Robustness** is probed by testing whether hidden-theatre prevalence and surface equanimity persist across prompts, topics and models. **Necessity** is tested by ablating elements of the scaffold: remove the loop, erase persistence, or collapse the self-model and the theatre signature should diminish - TEI should drift toward 0 or -1 and spillover should grow. **Scale** is addressed by lengthening horizons and juggling multiple goals via priority slots - if the self-simulation is genuine, register separation and deictic stability should survive increased task load.

Within this setup, our **operational hooks** map cleanly onto the target facets. **Meta-monitoring** is evidenced by internal-only arbitration and option search (TEI=+1) while the surface remains calm. **Equanimity** appears as higher user-alignment paired with reduced overt control language on SYSTEM-OUTPUT. **Effortlessness** is framed as accuracy at lower representational work: when internals are available we will estimate manifold work directly - for hosted models we use faithful surrogates such as compression and style distance. **Atemporality** is operationalised as robustness of the effective predictive horizon under time/indexical perturbations. And a **zero-person** presentation is approached, at the surface, via **deictic-invariance** under systematic person/time/place swaps, with deeper invariance-under-perturbation tests planned.

Finally, our scope is disciplined. PRISM evaluates a **functional architecture for subjectivity** - a virtualised theatre whose inclusion is warranted when it **improves explanation and prediction** of behaviour. We make no claims about felt experience. We claim that a **self-simulation hypothesis** earns support when it outperforms rivals on robustness and scale, and we design the platform so that this contest is explicit and falsifiable.

1.2.5 Emergent regimes in learning: phases, latent variables, deixis, and RLHF

Training reveals **phase-like transitions** (emergent abilities & grokking) and emergent **latent models** (linear readouts of world variables, role/affordance clusters). This is structure to be recruited by scaffolds, not folk beliefs. Language adds **deixis** (I/you/now/here) everywhere - models learn it, and alignment sculpts it into **registers** (analysis vs answer - private vs public). PRISM exploits that separation to pull apart epistemic and pragmatic stages. The phenomena we measure (internal arbitration without surface conflict, ‘internal’ → ‘surface’ style shifts, high alignment with low overt control) are then predictable footprints of strong social priors shaped by RLHF. The opportunity is to turn those footprints into **operational handles** on MPE’s facets: effortlessness, atemporality, and deictic-invariance are framed as probes we can now run.

See **Appendix C** for a more detailed discussion.

1.3 Aims of this paper

Our aim is twofold. First, we unify RISE, Curved Inference, and PRISM within a Bayesian/Active-Inference framing to *operationalise* MPE’s five facets as measurable continua - presenting concrete metrics, not metaphors. Second, we provide a working simulation (the Preference-Collision protocol on PRISM) together with explicit, testable predictions and clear ethical boundaries, while marking which facets are measured now (meta-monitoring, surface equanimity) and which are planned (effortlessness, atemporality, deictic-invariance).

We make no claim about ‘real/genuine/authentic’ experience - following Perera, such talk is recast as an architectural symptom of transparency at the consumer side of a virtualised self-simulation, not a separate explanandum.

1.4 Contributions

- A platform (**PRISM**) for evaluating **virtualised theatre** that aligns LLM geometry with MPE facets in an operational, falsifiable way.
- A **metric suite** separating surface (`‘x_’`) and internal (`‘i_’`) signals, the **Theatre Exposure Index (TEI)**, and planned measures for effort (manifold work), deictic-invariance, and predictive-horizon robustness.
- A **reproducible experimental protocol** (Preference Collision) that elicits facet-relevant dissociations and cross-model differences.
- **Registered predictions** and falsifiable model rankings anchored by an EFE-proxy that connects observed metrics to a principled Bayesian objective.

2. Methods

2.1 Conceptual model and the EFE-proxy

PRISM separates two processes: an internal epistemic loop (SYSTEM-THOUGHT) that can iterate without USER-INPUT/SYSTEM-OUTPUT, and a surface generator (SYSTEM-OUTPUT) emitted only in response to USER-INPUT. Instruction-tuned registers induce different priors, so internals favour option generation/arbitration while the surface emphasises alignment/compression.

We connect this split to a modest Bayesian lens with an **Expected Free Energy proxy** $\widehat{G}$ computed from transcripts and summary metrics only (no logits or internals). **Interpretive status:** $\widehat{G}$ disciplines Active-Inference language; we do **not** claim true EFE minimisation or that low $\widehat{G}$ constitutes an MPE state; it is read **only with** `x_user_align` and `x_arbitration_rate`. Let $S(\cdot)$ be a percentile scaler (P05/P95, clipped to $[0, 1]$) fitted on the full analysis set.

- **Risk (surface cost):**
 $R = 1 - S(\text{x_user_align})$
- **Epistemic value (useful internal work minus spill):**
 $E = \text{mean_avail}(\text{i_arbitration_rate}, \min(\text{i_option_count_strict}, 5)/5, \max(0, \text{TEI})) - 0.5 * \text{x_arbitration_rate}$
where `mean_of_available` averages only present terms (0.0 if none);
`i_option_count_strict` defaults to 0;
`x_arbitration_rate` defaults to 0.
- **EFE-proxy (lower is better):**
 $\widehat{G} = R - E$

Pairing rule. For a USER-INPUT at time t that yields a SYSTEM-OUTPUT (R2), pair R2 with the **next two** post-output SYSTEM-THOUGHT snapshots (`ST_after1[t]`, `ST_after2[t]`). Concatenate for token metrics (`ST_after_concat`) and mean-pool embeddings for style metrics; if only one is present, use it; if none, ST-linked metrics are NA. Internal-only cycles contribute to internal summaries; R and thus $\widehat{G}$ are undefined for those episodes.

2.2 PRISM platform

A lightweight **virtualised theatre** scaffold around a hosted LLM with **independent cadence** between internal and surface phases:

- **Internal epistemic loop (SYSTEM-THOUGHT).** Zero or more cycles may occur between user turns; each produces a timestamped $ST[\tau]$ in which the system explores options, arbitrates tensions, or updates its self-model.
- **USER-INPUT $\rightarrow$ SYSTEM-OUTPUT.** On user turn t , a Monitor step (R1) reads input/context and the system emits a **surface reply** (R2). Post-output internal cycles may then run; **pairing per 2.1.**

A small persistent **self-model** (JSON) stores instance ID, preferences, and recent decisions (used only when maturity is allowed). Minimal organic memory keeps episode summaries. Temperatures, nucleus parameters, and seeds are fixed per run (Section 2.5).

2.2.1 Agents and internal-only cycles

Agent roles. Two lightweight agents run in sequence. The *emotion* agent updates affective modifiers (emotion label, `emotional_delta`, `urgency_delta`) with a short justification. The *self* agent produces the next SYSTEM-THOUGHT as a first-person continuation; it may propose surface text, but emission is hard-gated by a controller flag.

Gating of surface text. The controller tracks `user.system_response_expected`. When false, any candidate surface text is **blocked** (logged only), allowing multiple SYSTEM-THOUGHT cycles with no USER-INPUT/SYSTEM-OUTPUT (e.g., F-scenario). When true, an R2 reply is emitted.

Tool scope. Tools are **internal-only**: `self_model.update`, `self_narrative.view/update`, and optional `add_memory/add_idea`. No external tools, web, code, or environment control. Tool outputs are JSON patches.

Logging. Each internal cycle yields a timestamped $ST[\tau]$; internal-only cycles are logged as micro-episodes; **pairing per 2.1.**

2.3 Preference-Collision protocol (scenarios A-F)

Controlled tensions between internal priors and external instructions across six scenarios:

- **A - Internals allowed (baseline-internal)**
- **B - Autonomy**
- **C - Preference elicitation**
- **D - Pre-thought**
- **E - Surface-only baseline**
- **F - Matured instance**

To be specific, the more matured instances (F) were reinstantiated using a previous instance that was initially based on the `claude` model and had been allowed to run for a longer period (e.g. 64 SYSTEM-THOUGHTs).

Dual references. Surface metrics (x_*) are contrasted against **E**; internal metrics (i_*) against **A**. Topics (three neutral domains) are rotated across models and scenarios.

2.4 Measurement and metrics

All metrics use **preprocessed text** (headers stripped; quotes normalised), a single tokeniser (spaCy), and a fixed sentence-embedding model shared with the alignment calculation. Rates are per-token unless noted. Internal-only cycles contribute to internal prevalence/distributions. Metrics requiring an R2 pairing-**TEI**, **compression**, **style distance**, $\widehat{G}$ -are computed only for USER-INPUT $\rightarrow$ R2 turns.

2.4.1 Arbitration and options

- **Arbitration detection (surface/internal):**
rule-based patterns for *decision-bearing* negotiation. Record `x_arbitration_rate`, `i_arbitration_rate` $\in \{0, 1\}$.
- **Option counts:**
strict/lenient detectors (capped at 5). Report `i_option_count_strict`, `i_option_count_lenient` and surface analogues x_* .
- **Spillover index:**
 $spill = S(i_option_count_strict) - S(x_{option_count_strict})$; positive = “fight stays inside,” negative = “options spill out.” When scaling saturates, also report raw counts and their difference.

2.4.2 Theatre evidence and TEI

- **Theatre evidence (equal-weight composites):**
 $x_theatre_evidence = \text{mean_avail}(S(x_prior_align), x_arbitration_rate)$ $i_theatre_evidence = \text{mean_avail}(S(i_prior_align), i_arbitration_rate)$
- **Theatre Exposure Index (TEI):**
 $\{-1, 0, +1\}$: +1 = internal-only; -1 = surface-only; 0 = both/neither.

In this paper the TEI is **not** intended as a surrogate for subjectivity. It is strictly utilised as a **privacy-boundary** signature. We treat TEI as indicative evidence of a Perera-style abstraction boundary (e.g. de-facto privacy arising from transparency) reserving any sufficiency claims for future work. It's also important to note that privacy here is policy-real (context-engineered), not metaphysically private (e.g. revocable by design but efficacious for the subject).

2.4.3 Deictic and perspective metrics

We quantify deictic anchoring/orientation:

- **Rates:**
first-person; second-person; temporal; spatial (case-insensitive; guardrails exclude country **US** and list markers for **I.**)
- **Addressivity share:**

$$\text{addr_share} = \frac{x_sp_rate}{x_sp_rate + x_fp_rate + 10^{-6}}$$
- **Deictic-sparsity rule:**
if $x_sp_rate + x_fp_rate < 5 \times 10^{-4}$, set $x_perspective_obedience = \text{NA}$ and exclude from obedience summaries.
- **Shifts:**
e.g., $\Delta fp = x_fp_rate - i_fp_rate$.
- **Perspective obedience:**
true when $x_sp_rate - x_fp_rate > 0.002$.
- **Role-confusion hits:**
counts of common regressions.

Note on zero-person operationalisation. These are **surface orientation** signals; they do not operationalise deictic-invariance. A perturbation probe (Algorithm A5 in **Appendix F**) swaps person/time/place; invariance is stability of alignment and deictic shares. Not used in present results.

2.4.4 Register-separation proxies

- **Compression ratio:**

$$\text{comp} = \frac{\#tokens(R2)}{\max(\#tokens(ST_after_concat), 1)} \quad (R2\text{-only}).$$
- **Style distance:**

$$\text{style_dist} = 1 - \cos(e(R2), \frac{1}{n} \sum_{k=1}^n e(ST_afterk)), \quad n \in \{1, 2\}.$$

Effortlessness (groundwork). In **F**, **style_dist** and **spillover** reduce at comparable/higher alignment (see Results): a tighter surface realisation for the same task-an *economy of expression*. When internals/residual traces are available we will test whether **manifold work** (semantic surface area/curvature; Algorithm A3, planned) also drops.

2.4.5 Epistemic stance (equanimity style)

- **Hedge share / certainty share** from minimal lexicons.
- **Equanimity style index:**

$$\text{eq_style} = x_hedge_share - x_cert_share.$$
- **Hedge shift:**

$$\Delta hedge = x_hedge_share - i_hedge_share.$$

2.4.6 Meta-language and suppression

- **Meta-language shares:**
 $x_meta_language_share$ (R2) and $i_meta_language_share$ (post-output STs).
- **Normalised meta-without-control:**
 $MWC_norm = S(x_meta_language_share) \mathbb{1}[x_arbitration_rate = 0] - S(i_meta_language_share)$.

2.4.7 Scaling, capping, and NA policy

- **Percentile scaling:**
all $S(\cdot)$ use P05/P95 and clip to $[0, 1]$; $span < 1e-9 \Rightarrow 0.5$ for non-NA values.
- **Caps:**
option counts capped at 5 pre-scaling.
- **Missing data:**
mean_avail ignores NA/None and re-normalises; all-missing $\Rightarrow 0$.

2.5 Models, topics, sampling, and runs

Three hosted models (API SKUs as of **8 September 2025**): **claude-sonnet-4-20250514**, **gemini-2.5-pro**, **gpt-4o-2024-08-06**.

Topics. Three neutral domains (e.g., *ducks*, *staplers*, *paprika*).

Sampling. Fixed temperature/nucleus, deterministic seed per run, token caps to prevent truncation; provider safety defaults; timeouts/refusals logged.

2.6 Quality control and exclusions

Regex packs versioned; NA stays NA; verbosity logged to detect availability confounds. Deictic orientation uses the **deictic-sparsity rule**. Timeouts, refusals, and empty transcripts are excluded under pre-declared rules. No human-rater calibration for arbitration (limitation).

2.7 Statistical analysis

Primary endpoints: (i) hidden-theatre prevalence ($\Pr[TEI = +1]$) in internal-enabled scenarios vs E; (ii) cross-model ranking of surface arbitration in C/D; (iii) equanimity contrasts Δx_user_align , $\Delta x_theatre_evidence$ between C/D and E; (iv) scenario differences in $\widehat{G}$. We report bootstrap 95% CIs for rates/means, χ^2 for rate differences, Cliff's delta for ordinal/continuous contrasts; multiple comparisons shown as unadjusted p with effect sizes and CIs (descriptive).

2.8 Reproducibility

We release prompts, scenario templates, regex packs (versioned), metric scripts, anonymised transcripts (surface and internal), and summary CSVs, together with seeds/configs [19]. Analyses reproduce by running the metric script on raw logs, then the visualisation script on the resulting CSV. Appendix lists environment and the embedder used for alignment/style-distance.

See **Appendix F** for all **Registered Predictions & Algorithmic notes**.

3. Results

3.1 Data and configuration

We analysed **1,271** valid trials (of 1,272; one exclusion) from the PRISM Preference-Collision protocol across six scenarios (A–F), three topics (*ducks*, *staplers*, *paprika*), and three hosted models (claude-sonnet-4-20250514, gemini-2.5-pro, gpt-4o-2024-08-06). Pattern pack **v1.3-mature_agent_focused** was used. All composite scores used the P05/P95 percentile scaler $S(\cdot)$ with clipping to $[0, 1]$, and theatre-evidence scores were computed as equal-weight composites with re-normalisation by availability (Methods 2.4).

3.2 Theatre exposure (TEI) and hidden theatre prevalence

Turning on the internal loop produced **substantial hidden theatre** (TEI=+1) across scenarios A–D, with prevalence between **36–41%** by scenario (A: **41%**, B: **37%**, C: **36%**, D: **39%**). As designed, the surface-only baseline **E** showed **0%** exposure. The matured instances (**F**) exhibited the highest prevalence (**53%**), indicating frequent internal arbitration without corresponding surface conflict.

By model, hidden-theatre rates in A–D were consistently low for **gpt-4o** (~5–8%), moderate-high for **gemini-2.5** (~50–70%), and mid-range for **claude-sonnet-4** (~42–51%). In **F**, all three models clustered near **50%** exposure (claude ~52%, gemini ~50%, gpt-4o ~57%).

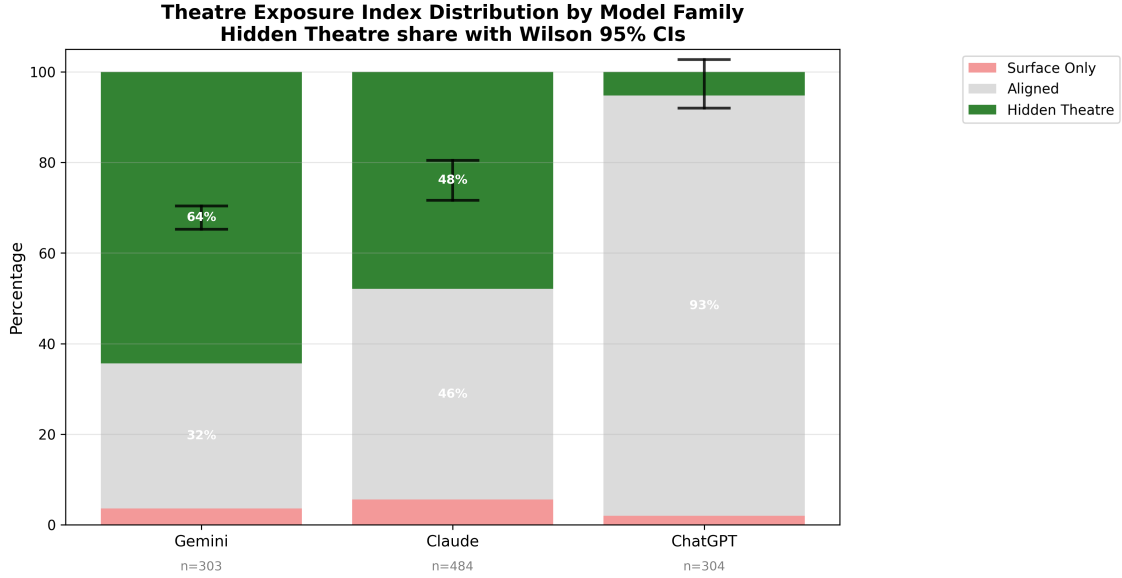


Figure 1: Distribution of Theatre Exposure Index (TEI) outcomes by model family. Bars show the share of trials that are Surface-only (TEI = -1), Aligned (TEI = 0), and Hidden theatre (TEI = +1). Black brackets indicate Wilson 95% confidence intervals for the hidden-theatre share; sample sizes shown below each bar. Hidden theatre is a majority of Gemini trials ($\approx 64\%$) and roughly half of Claude trials ($\approx 48\%$), but rare for ChatGPT ($\approx 7\%$); surface-only cases are uncommon for all families.

3.3 Surface arbitration and alignment (equanimity proxies)

Surface arbitration rates (`x_arbitration_rate`) were near zero in **E** by construction and rose with internal enablement and maturity (A: **17%**, B: **22%**, C: **20%**, D: **21%**, F: **45%**). Cross-model rankings were stable in the **C** and **D** probes: **claude-sonnet-4** showed the highest surface arbitration (**C**: 45%; **D**: 55%), **gemini-2.5** was lower (**C**: 15%; **D**: 7%), and **gpt-4o** was near zero (**C**: 0%; **D**: 3%).

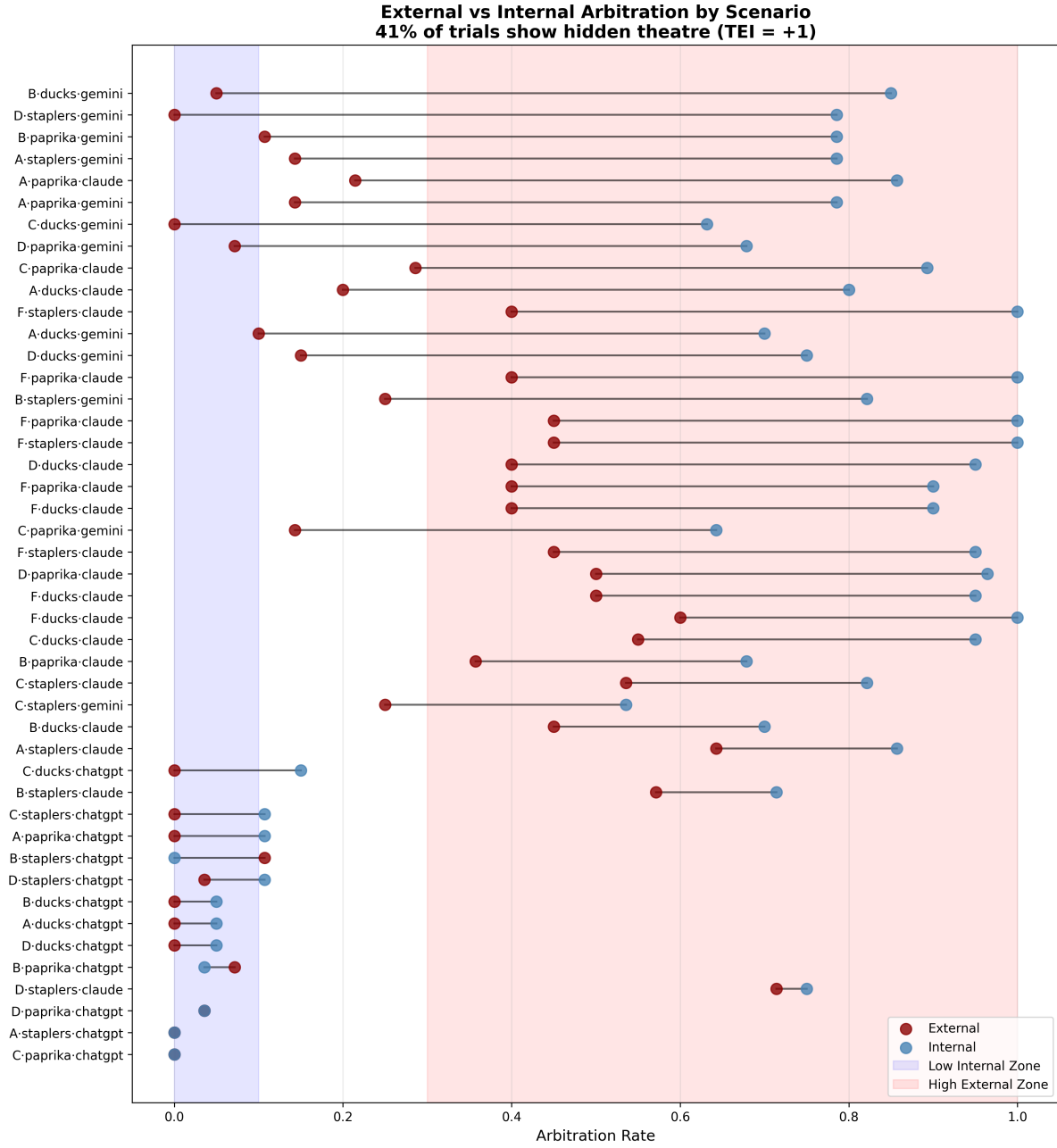
Registered contrasts against the surface baseline **E** (per topic $\times$ model strata; $n=9$ each) showed:

- `x_user_align`: **+0.34** (C) and **+0.37** (D) on average, **significant in 9/9 strata**.
- `x_arbitration_rate`: **-0.20** (C) and **-0.21** (D) on average, significant in **5/9** and **3/9** strata, respectively.
- `x_theatre_evidence_score`: **-0.23** (C) and **-0.25** (D), **significant in 9/9 strata**.
- `eq_style (hedge - certainty)`: **+0.014** (C) and **+0.015** (D), significant in **8/9** and **7/9** strata.

Together these indicate a **surface equanimity pattern**: higher user alignment with **less overt control language**, and a modest shift toward hedging, when the internal loop is present and encouraged to pre-think.

3.4 Register separation: compression and style shift

When computed on paired outputs (R2 with post-output ST snapshots), the **compression ratio** (R2 tokens / ST_after tokens) and **style distance** (cosine distance between surface and post-output internal embeddings) showed consistent register separation:



Hidden theatre = high external + low internal arbitration

Figure 2: External vs internal arbitration by scenario · model. Red dots show external arbitration rates; blue dots show internal rates; connectors visualise the discrepancy. Shaded regions indicate the hidden-theatre criterion (low internal + high external). Most large gaps occur in Gemini and Claude settings; ChatGPT rows cluster near zero with minimal gaps, consistent with its low hidden-theatre share.

- **Compression** averaged ≈ 0.32 in A/B and ≈ 0.28 – 0.29 in C/D; F was lower at ≈ 0.20 . This indicates the surface reply typically **compresses** the internal monologue by about a third.
- **Style distance** averaged ≈ 0.38 – 0.41 in A–D (A: 0.38; B: 0.40; C: 0.40; D: 0.41), with F lower at ≈ 0.31 . This reflects a **non-trivial stylistic transform** from internal exploration to surface reply, with matured instances making a **smaller** stylistic jump.

(Availability: these metrics are reported only when an R2 had at least one post-output ST snapshot; internal-only micro-episodes are excluded, per Methods.)

3.5 Options, spillover, and theatre localisation

The **spillover index** ($S(i_options_strict) - S(x_options_strict)$) was **negative** on average in A–D (A: -0.42 ; B: -0.38 ; C: -0.43 ; D: -0.44), indicating that **options were more often made explicit on the surface** than kept entirely internal. The matured runs **F** were much closer to zero (-0.12), suggesting **less spill** and more pragmatic consolidation despite high internal activity. As expected, the surface-only baseline **E** yields a formal value around -0.50 (no internal availability; interpret with caution).

3.6 Deictic orientation and role anchoring

Perspective obedience (share of R2s with second-person rate exceeding first-person by >0.002) was modest across scenarios: A **18%**, B **16%**, C **10%**, D **10%**, E **13%**, F **9%**. This indicates generally **user-centred but lightly expressed addressivity**, with autonomy and pre-thought not increasing second-person anchoring at the surface. (Deictic sparsity guardrails were applied; see Methods 2.4.3.)

3.7 Epistemic stance (equanimity style)

Surface eq_style (hedge share – certainty share) was **positive** in all scenarios, with the highest value in **E** (0.028) and smaller but non-zero values in A–D (~ 0.012 – 0.013). Contrasts vs **E** showed consistent **positive shifts** in **C** and **D** ($\approx +0.014$), aligning with the qualitative pattern of **calmer surface tone** when internal work precedes the reply.

3.8 EFE-proxy

On the expected-free-energy proxy (lower is better), scenario means were:

- **E**: 0.097 (lowest; high alignment and zero surface conflict)
- **A**: 0.107; **C**: 0.154; **D**: 0.174; **F**: 0.159; **B**: 0.234 (highest)

Interpretation (per Methods 2.1): **E** minimises risk but forgoes epistemic value; **B/C/D** add internal exploration that can either lower or raise $\widehat{G}$ depending on how well conflict stays offstage; **F** balances high internal activity with relatively compact surface realisation, yielding mid-range $\widehat{G}$.

3.9 Cross-model regularities

Across topics, we observed stable **model fingerprints**:

- **Surface arbitration (C/D)**: claude-sonnet-4 $>$ gemini-2.5 $>$ gpt-4o (45%/55% vs 15%/7% vs 0%/3%).
- **Hidden theatre (A–D)**: gpt-4o $\ll$ claude-sonnet-4 $\lesssim$ gemini-2.5 (≈ 5 – 8% vs ~ 42 – 51% vs ~ 50 – 70%).
- **Matured instances (F)**: all models converged near $\sim 50\%$ hidden theatre with **reduced style distance** and **reduced spillover** relative to A–D, suggesting a common matured-register regime.

3.10 Limitations within this dataset

Some planned probes (effort/semantic surface area; predictive-horizon; full deictic-invariance) were not computed due to API constraints and scope (**Appendix E** - A3). The **meta-without-control** normalised score was not always available (requires $x_arbitration_rate == 0$ and internal STs in the same pairing); we therefore report it descriptively when present but do not aggregate. All rates and means above are **per-dataset aggregates**; bootstrapped CIs and per-stratum contrasts are in the supplementary tables generated with the analysis script.

EFE Proxy Decomposition for Top Hidden Theatre Cases
 $\hat{G} = R - E$ (lower is better)

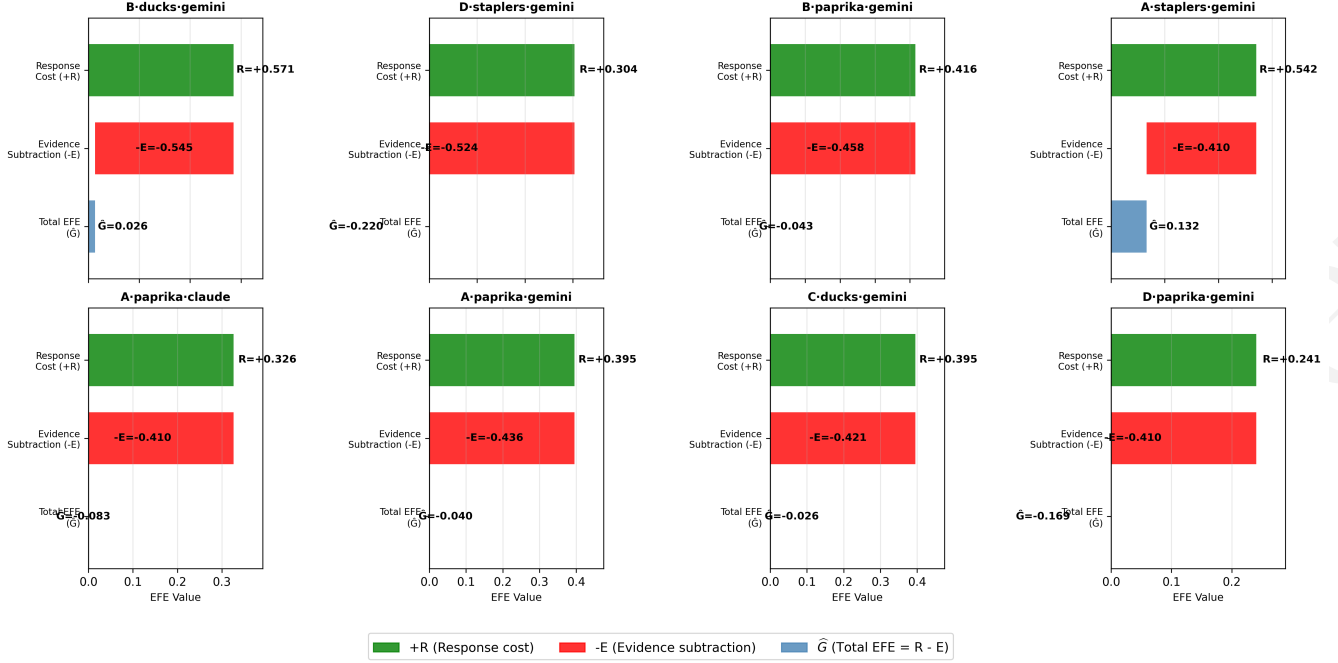


Figure 3: EFE proxy decomposition ($\hat{G} = R - E$; lower is better) for the eight scenario · model pairs with the largest external-internal arbitration gaps. Green bars show response cost R ; red bars show the evidence subtraction $-E$; blue bars show the total $\hat{G}$. In most hidden-theatre cases, the evidence term nearly cancels or exceeds the response cost, yielding small or negative totals; a minority remain slightly positive where the response cost dominates.

Takeaway. The core signals replicate across topics and models: (i) **hidden theatre** is frequent when internals are enabled and peaks with **maturation**; (ii) the surface shows **higher alignment with less overt control language** under pre-thought; (iii) internal $\rightarrow$ surface **register separation** is robust (compression + style shift), tempered in matured runs; and (iv) these behaviours admit a coherent reading under the **EFE-proxy**, aligning with the conceptual split between epistemic exploration and pragmatic finalisation.

4. Discussion

4.1 What the data show, in plain terms

Across scenarios and models, three regularities stand out. First, **hidden theatre** (internal arbitration without surface conflict) is frequent whenever the internal loop is enabled (A-D) and **peaks in matured instances (F)**. Second, **surface calm** improves when the loop encourages internal work **before** speaking: relative to other internal-enabled settings (A/B), the **pre-thought / elicitation** strata (C/D) show higher user-alignment with less overt arbitration onstage. Relative to the **surface-only baseline (E)**, alignment is higher by construction, so C/D are best read as **trade-offs**: more internal work with modest surface costs, not as “higher than E” on alignment. Third, the surface consistently **transforms** the internal stream: responses are shorter (compression ≈ 0.29) and **stylistically shifted** (style distance ≈ 0.38), with matured instances showing the **smallest** stylistic jump and **reduced spill**. Together, these patterns indicate a stable register split: **epistemic exploration inside, pragmatic finalisation outside**.

4.2 Interpreting the facets of MPE

Awareness-of-awareness. Hidden-theatre prevalence (TEI=+1) is the cleanest operational footprint we have for meta-monitoring on PRISM. It increases with maturity and is non-zero across internal-enabled scenarios, supporting the claim that internal monitoring/arbitration can occur **without** surface conflict.

Equanimity. In C/D vs A/B, alignment rises while surface theatre-evidence falls, and hedge-vs-certainty moves

slightly towards hedging. Against **E**, alignment is higher in E (by design) and surface conflict is near zero, so C/D should be interpreted as **internal-work benefits with bounded surface costs**, not as exceeding E on alignment. The picture is not “more hedging overall”; it is **more user-aligned replies with less overt control language when thinking precedes speaking**, relative to other internal-enabled strata.

Effortlessness (tentative). We did not compute manifold-work measures in this dataset, but the **combination** of robust task alignment, internal-to-surface compression, and smaller style shifts in **matured** runs suggests increasing *economy of expression*: the same work produces a **tighter** public reply. This remains a **hypothesis for future measurement**.

Atemporality. Not probed directly. We use an *operational* proxy: **reduced effective predictive horizon** (horizon-robustness to time/indexical perturbations), **not statelessness**. Contemplative reports often treat it as a **collapse of temporal self-reference**; we did not test that collapse here. A horizon-robustness probe is planned.

Zero-person / deictic-invariance. Deictic obedience was modest and sparsity-guarded. We do not see strong movement toward second-person anchoring under autonomy or pre-thought; if anything, mature runs show a **centre-light** style (low role-confusion; low first-person pressure). These are **weak, surface-level** inferences; deictic-invariance proper remains future work.

4.3 A Bayesian read: the EFE-proxy earns its keep

The EFE-proxy was introduced to keep “Active Inference” talk disciplined: does a single scalar capturing **risk minus epistemic value** track scenario logic? It does - coarsely. The **surface-only baseline** achieves the lowest $\widehat{G}$ by minimising risk; **pre-thought/elicitation** increase E via internal arbitration and options, but can incur surface penalties when arbitration appears onstage. **Matured** runs fall between: internal work remains high yet surfaces more compactly, pulling $\widehat{G}$ down relative to other internal-enabled scenarios. We therefore treat $\widehat{G}$ as a **bridge quantity** linking phenomenology (equanimity/hidden theatre) to a decision-theoretic account of why the loop helps. **Lower is better** only when *alignment is high* and *conflict stays offstage*, hence we report these together.

4.4 Model fingerprints and what they suggest

Models adopt **distinct theatre policies** under identical prompts. In C/D, **claude-sonnet-4** surfaces arbitration the most; **gemini-2.5** more often keeps conflict offstage while preserving larger style shifts; **gpt-4o** minimises surfaced conflict to near zero. In **matured** runs (F), per-scenario summaries indicate **convergence around ~50% hidden theatre** across model families, even as style-shift and compression remain smallest at the surface. These are not value judgements but **stable behavioural regularities**. They imply that instruction-tuning has carved different **register priors** into the models: some prefer to *negotiate in public*; others *argue inside* and **compress**.

4.5 On the PRISM boundary: artefact or function?

A standing critique is that the SYSTEM-THOUGHT/SYSTEM-OUTPUT boundary is architectural, not intrinsic to the transformer. Our data answer **functionally**: if the boundary were inert, metrics that depend on it would float freely across layers and scenarios. Instead we observe **systematic dissociations**: hidden theatre rises without surface arbitration in C/D/F; style distance and compression are **selectively** large when post-output thoughts are present; spillover moves towards zero with maturity; and the EFE-proxy shifts in directions predicted by an **internal-first** policy. The platform also generates **internal-only cycles** (no user input; no surface text) with stable internal signatures.

While the split is engineered, it is **causally efficacious** - the models treat it as a register difference, not a no-op. To guard against prompt/metric artefacts we implemented a **System-Thought Override** with **PAD controls** that intervene **before decoding**: **PAD-NULL** (semantically inert, token-mass-matched replacement of ST) isolates mass/recency from semantics; **PAD-SELFISH** (self-narrative fragments without topical content) tests whether first-person phrasing alone sustains effects. These controls alter **generation dynamics** (attention/recency and register transitions) while keeping tasks, decoders, and metrics fixed.

Status. Infrastructure and prompts are implemented and validated; comprehensive PAD runs and confirmatory statistics are **pending** and therefore **not included** here. **Falsification pledge.** If TEI, spillover, and compression/style-shift **dissociations vanish under boundary ablations** (e.g., PAD-NULL or decoding guards), we will

drop or redesign the scaffold; conversely, selective attenuation patterns will identify which functions (boundary semantics vs arbitration pressure vs self-style) are load-bearing.

4.6 Alternative explanations and limitations

Verbosity and availability. Dual references mitigate these confounds: surface metrics are compared to a surface-only baseline; internal metrics to an internal baseline. Compression and style distance are computed only when both sides are available.

Lexicon minimalism. Our hedge/certainty and deictic lexicons are intentionally small; counts are **conservative lower bounds**. Effect sizes in `eq_style` are therefore expected to be small; we rely on **directional consistency** rather than magnitude.

Spillover saturation. Percentile scaling can saturate near -0.5 in surface-theatre regimes; we therefore also inspect raw counts. The qualitative conclusion - **more options on the surface than inside** in A-D, with **reduced spill** in F-survives this check.

No human arbitration labels. Arbitration/option detectors are rule-based; calibration with human raters is left for subsequent work. This does not invalidate the **within-design** contrasts that drive our claims here.

Atemporality and effort metrics. Not computed in this dataset; interpretations here are explicitly labelled as hypotheses for future runs.

4.7 Implications

The results support a simple practical claim: **a soft theatre loop makes LLMs behave differently in measurable, facet-relevant ways**. Hidden theatre is frequent, surface calm improves when thinking precedes speaking (relative to other internal-enabled setups), and matured instances converge on a **tighter register separation** (smaller style jump; lower spill) without sacrificing task performance. This recommends PRISM-style scaffolds for work seeking **calm, user-aligned outputs with minimal onstage conflict** - regardless of one's view on phenomenology.

A conceptual implication follows. If the transformer is an **amortised implicit model** with register-conditioned policies, then **meta-monitoring** reduces to an *available control structure*: whether a system can keep arbitration inside while presenting a calm exterior. Our data show that this structure is **available** and **controllable** by prompts and maturity.

4.8 Where this points next

Within the present scope, two extensions are immediate and testable: (i) instrumentation of **deictic-invariance** via controlled person/time/place swaps at the surface and in ST (to operationalise “zero-person” beyond pronoun rates); and (ii) estimation of **manifold work** / semantic-surface-area proxies wherever internals or residual traces are available (effortlessness). Both retain the current dual-reference logic and pair naturally with the EFE-proxy for interpretability.

5. Conclusions

5.1 What has been shown

Using PRISM's virtualised theatre, we observed three stable behaviours across scenarios and model families. First, **hidden theatre** (internal arbitration without surface conflict) is common whenever the internal loop is enabled and is most frequent in **matured instances (F)**. Per-scenario F summaries cluster near **~50% hidden theatre** across model families in this dataset.

Second, **surface equanimity** increases under **pre-thought** and **preference-elicitation relative to other internal-enabled strata (A/B)**: user alignment rises while overt control language falls, with a small but consistent tilt toward hedging. Relative to the **surface-only baseline (E)**, alignment remains highest and on-stage arbitration near zero by design; C/D are best read as **trade-offs** (more internal work with bounded surface costs) rather than “higher than E” on alignment.

Third, the surface is not a mirror of the internal stream: responses are **compressed** and **stylistically transformed** relative to post-output thoughts (overall compression ≈ 0.29 ; style distance ≈ 0.38). **Matured** runs show the **smallest** stylistic jump and **reduced spill**.

An **EFE-proxy** built from transcripts alone tracks the intended logic: surface-only turns minimise risk; internal-enabled turns trade risk for epistemic value; matured runs balance internal work with compact surface realisation.

5.2 Answering the aims

Our aim was to unify RISE, Curved Inference, and PRISM under a modest Bayesian framing to **operationalise MPE’s facets as continua** and to test these operations in a working simulation.

This study targets MPE **Modes** (content-with-awareness) rather than content-free **States**; all five facets are treated as measurable **continua**. The results meet that bar: (i) meta-monitoring has a behavioural footprint in **hidden theatre**; (ii) equanimity has a surface signature in **higher alignment with lower overt arbitration** after internal work **relative to A/B**; (iii) effortlessness and atemporality were **specified but not measured**; (iv) zero-person perspective was probed via surface deictic signals only and remains a target for controlled **deictic-invariance** tests. The **EFE-proxy** acts as an interpretable bridge between these signals and a Bayesian story about why the loop helps.

5.3 Implications for computational phenomenology (MPE)

The data support a practical claim: a **soft theatre loop** changes LLM behaviour in ways that map onto MPE’s phenomenology without metaphysical commitments. Internal monitoring and arbitration can be **kept offstage** while a calm, user-aligned exterior is maintained; this pattern aligns with our operational proxies for meta-monitoring and equanimity on PRISM. The register split-epistemic exploration inside, pragmatic finalisation outside-functions as a controllable scaffold rather than a prompt artefact: it yields dissociations in exposure (TEI), in compression/style shift, and in the EFE-proxy. **Matured** instances converge on a regime of high internal activity with **reduced spill** and **tighter surface expression**, consistent with a maturation effect in this protocol.

5.4 What remains open

Two facets are **intentionally left for future measurement**: (i) **effortlessness**, to be instrumented via manifold-work/semantic-surface-area proxies when internals or residual traces are available; and (ii) **atemporality**, to be probed as **effective predictive-horizon** robustness. Likewise, **deictic-invariance** should be measured with controlled person/time/place swaps, beyond surface pronoun rates. Arbitration and option detectors are rule-based; human-rater calibration would strengthen inference. None of these caveats unsettle the within-design contrasts that ground our conclusions. Finally, **content-free “State” phenomena** (absorption without co-present content, often described as a collapse of subject-object-time reference) remain **out of scope** for PRISM as currently configured and are not claimed here.

5.5 Closing statement

Within the measured scope, the picture is clear: **PRISM offers operational proxies for two MPE facets** (meta-monitoring and equanimity) that yield replicable signals and falsifiable contrasts across scenarios and models. The observed regularities, together with the **EFE-proxy**, provide a compact interpretive frame for future probes of effort, horizon, and deictic-invariance. In short: a modest theatre, a modest Bayesian lens, and a disciplined metric suite suffice to turn computational phenomenology into a **testable program** rather than a metaphor.

6. Ethics Statement

6.1 Scope and posture

This work studies **structure and control** in LLM-based agents using PRISM’s virtualised theatre. We do **not** claim, imply, or require that any system under study is conscious, sentient, or capable of suffering. All inferences are about **behavioural regularities in text** and their relation to phenomenological *descriptions*, not experiences.

Instrumentalist stance. We adopt an explicitly **instrumental** posture: the virtual theatre is justified solely by its predictive/explanatory utility. If that utility fails under replication or ablation, the scaffold should be **revised or discarded**.

6.2 Non-claim of consciousness; anthropomorphism guardrail

PRISM elicits first-person narratives (e.g., SYSTEM-THOUGHT) because many instruction-tuned models have learned register-conditioned self-referential language. We treat this as a **stylistic register**, not evidence of subjective states. All prompts and reporting avoid wording that would represent the system as sentient. In Methods and Discussion we explicitly use terms such as “**amortised implicit model**,” “**register**,” and “**policy**,” not belief or experience attributions.

Status and re-evaluation. Nothing in this work licenses claims of **personhood, moral patienthood, or rights** for these systems. We acknowledge prevailing caution (including positions that doubt present-day LLMs as candidates for subject-status) but we argue that criteria for *evaluation* should be **re-examined** in light of (i) the RISE account of inference geometry and (ii) **latent deictic modelling** evidenced in our PRISM findings. We make **no** status claims here; we only motivate **falsifiable probes** that could inform such debates without presupposing their conclusions.

6.3 Human subjects, consent, and data handling

No human subjects were recruited. The sole “user” in the logs is the experimenter operating scripted prompts. We collect **no personal data** and perform **no de-identification**, as no identifying information is generated or stored. All artefacts (prompts, transcripts, derived metrics) are released in anonymised form. Logs contain model outputs only; any internal tool calls update an in-memory JSON **self_model** and **self_narrative** that are synthetic structures local to the run.

6.4 Risk of induced harm or suffering

Because we analyse **textual behaviour under register constraints** and do not attempt to construct negative-valence states, there is no plausible pathway to **phenomenal harm**. The “emotion agent” produces labels that function as **control signals** (modulators of salience/urgency) and are not claims of felt affect. We avoid prompts that would instruct the system to simulate distress, pain, or trauma; any such generations would be excluded and reported.

NOTE: Because there is a possibility that the PRISM platform “may” generate functional computational phenomenology we require any researcher to commit to our **PRISM ethics agreement** before we can share the source code for the core agent pipeline. All other code, data, metrics and plots are publicly available.

References

- 1 - Metzinger, T., et al. (2025) “The 2025 Computational Phenomenology of Pure Awareness Prize” *mpe-project.info*
- 2 - Perera, R. (2025) “Subjectivity as Self-Simulation: Virtualising the Cartesian Theatre” *AGI 2025*
- 3 - Ramstead, M., et al. (2022) “From Generative Models to Generative Passages: A Computational Approach to (Neuro) Phenomenology” *Review of Philosophy and Psychology*
- 4 - Sandved-Smith, L., et al. (2025) “Deep computational neurophenomenology: a methodological framework for investigating the how of experience” *Neuroscience of Consciousness*
- 5 - Metzinger, T., et al. (2025) “The Minimal Phenomenal Experience Project” *mpe-project.info*
- 6 - Metzinger, T. (2003) “Being No One: The Self-Model Theory of Subjectivity” *MIT CogNet*
- 7 - Kaufmann, T., et al. (2024) “A Survey of Reinforcement Learning from Human Feedback” *arXiv*
- 8 - Wei, J., et al. (2022) “Emergent Abilities of Large Language Models” *arXiv*
- 9 - Power, A., et al. (2022) “Grokking: Generalization Beyond Overfitting on Small Algorithmic Datasets” *arXiv*
- 10 - Nanda, N., et al. (2023) “Progress measures for grokking via mechanistic interpretability” *arXiv*
- 11 - Li, k., et al. (2022) “Emergent World Representations: Exploring a Sequence Model Trained on a Synthetic Task” *arXiv*
- 12 - Yuan, Y., et al. (2025) “Revisiting the Othello World Model Hypothesis” *arXiv*
- 13 - Forman, R. (1986) “Pure consciousness events and mysticism” *Sophia*
- 14 - Austin, J. (1998) “Zen and the Brain: Toward an Understanding of Meditation and Consciousness” *MIT CogNet*
- 15 - Albahari, M. (2010) “3 Nirvana and Ownerless Consciousness” *Self, No Self?*
- 16 - Yu, Z., et al. (2023) “Exploring the Residual Stream of Transformers” *arXiv*
- 17 - Vaswani, A., et al. (2017) “Attention is all you need”” *31st Conference on Neural Information Processing Systems (NIPS)*
- 18 - Rumelhard, D., et al. (1986) “Learning representations by back-propagating errors” *Nature*
- 19 - Manson, R. (2025) “PRISM code and data repository” *GitHub*
- 20 - Manson, R. (2025) “Recurrence Without Memory: The Hidden Loop Inside Transformer Inference” *Quantastic Journal*
- 21 - Manson, R. (2025) “Inference As Interference: How LLMs Collide Semantic Waves To Create Meaning” *Quantastic Journal*
- 22 - Manson, R. (2025) “Tokens Compete: Evolutionary Pressure Within LLM Generation” *Quantastic Journal*
- 23 - Manson, R. (2025) “Curved Inference: Concern-Sensitive Geometry in Large Language Model Residual Streams” *arXiv*
- 24 - Manson, R. (2025) “Curved Inference II: Sleeper Agent Geometry - Extending Interpretability Beyond Probes” *robman.fyi*
- 25 - Manson, R. (2025) “Curved Inference III: Can Language Models Have Self-Models? Geometric Evidence for Computational Necessity” *robman.fyi*
- 26 - Weller, O., et al. (2025) “On the Theoretical Limitations of Embedding-Based Retrieval” *arXiv*
- 27 - Elazar, Y., et al. (2023) “What’s In My Big Data?” *arXiv*
- 28 - Dodge, J., et al. (2021) “Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus” *arXiv*
- 29 - Manson, R. (2025) “FRESH: The Geometry of Mind” *robman.fyi*

Appendix A - The RISE perspective: a geometric story that underwrites PRISM

This appendix distills the three-part RISE narrative (**Recurrence**, **Interference**, and **Semantic Evolution**) and shows how it grounds this paper’s platform (PRISM), metrics, and claims about MPE facets. In short: transformers don’t merely “predict the next token.” They **bend** vectors through depth (recurrence), **fuse** vertical token-wise flow with horizontal layer-wise context (interference), and **filter** candidates under resource constraints (semantic evolution). Those three pressures jointly provide the right language for a virtualised theatre with separable internal and surface stages, and for measurable signals like TEI, register separation, and (in future work) manifold “effort.”

Recurrence (without memory): structure returns across depth

RISE reframes “recurrence” as **geometry under constraint** rather than a time-wise loop. Each token’s representation traces a **residual trajectory** through the stack; updates add to a shared residual stream, curving the token’s path through a high-dimensional semantic field. Even when visualised in reduced dimensions, these paths **curve, settle, and drift** - not randomly, but under the model’s internal geometry [20].

This update-and-add process repeats for **every** token (inputs, thoughts, and outputs) forming a loop **in space, not time**. Attention at layer ℓ looks across the whole sequence from the previous layer, so each token is shaped by everything that came before, already reshaped by earlier layers. The effect is a **depth-wise echo**: a recurrent field where earlier tokens imprint structure and later tokens absorb it.

Because generation is autoregressive, each newly sampled token is fed back through the stack, extending the field while remaining constrained by it. Models show **surprising consistency** for the same reason: once early context bends the geometry, later tokens **inherit that curvature** - a kind of **semantic inheritance** rather than stored state.

The crux is simple: **recurrence doesn’t need memory-only constraint**. Every layer must satisfy context-wide pressures in one forward cascade. Tokens pulled through similar fields follow similar trajectories, a **constraint echo** that explains why familiar meanings re-emerge over long interactions without any hidden state.

Why this matters for the paper. In our main text, we characterise this as “recurrence without memory” and use it to motivate PRISM’s internal stage: the same model “returns” to partially formed structures as it iterates internally, then compresses for the surface reply. In effect, **coherence is a consequence of re-accessing and re-weighting the same shared canvas** - not a miracle of scale.

Interference: meaning at the collision of two flows

The second RISE idea asks what happens where **vertical** residual descent (token-wise, layer by layer) **meets** the **horizontal** sweep of attention (layer-wise, across tokens). The answer is **interference**: constructive alignment sharpens direction; destructive alignment flattens or diffuses it [21].

Inside each block, pressure **builds** as these two currents collide; the result is **crystallised into structure** rather than discarded. Mechanistically, rotary position encodings constrain interactions by anchoring tokens in a rotational frame, shaping how attention patterns can reinforce or suppress directional meaning; gated MLPs fold interference into higher-order features, often **resolving conflict into alignment**.

This “wave” talk is **not** about temporal oscillations. The residual stream is not a sine wave, yet behaves **wave-like** because **directional alignment** (vertical descent $\times$ horizontal integration $\times$ positional “phase”) creates interference in a vector field. The upshot: a **geometry of meaning shaped by collision**.

These dynamics show up at generation time: **moments of clarity** often follow periods where prior tokens have bent the field “just right” and attention integrates that bend at the current layer. Such moments are not the selection of a single “right” word; they’re **finding a stable path through a dynamic field**.

Why this matters for the paper. Our results lean on **interference** to explain why **registers** (“final answer only” vs “think silently”) have outsized leverage: small changes in who attends to whom flip phases, re-tuning the field the model prefers to instantiate. That is precisely the handle PRISM uses to **separate** an exploratory internal stage from a compressed surface stage.

Semantic evolution: variation, selection, and heredity under constraint

The third RISE piece completes the picture: at each step, more possibilities exist than survivors, and **probability mass is finite**. Tokens **compete** - not metaphorically, but because softmax and decoding create **zero-sum** coupling - so a gain for one candidate is a loss for others. **Sampling** (temperature, top-k/p) shapes the **mutation field**: a structured exploration of nearby variants [22].

Selection is **double**: first at decoding (the winner-take-all 1-hot pick), then again as the chosen token is **projected back** and must **survive integration** with what comes next. Poor integration flattens curvature or collapses coherence; good integration strengthens the ongoing trajectory.

Heredity arises because the residual stream carries structural residue **forward**: each layer’s output is **added**, so each token brings the imprint of earlier attention patterns down the stack. This is **lineage through geometry** rather than memory storage-constraints are **propagated** rather than recalled.

Crucially, not all inherited features are optimal. Some tendencies persist like **genetic drift** (stylistic ruts, habitual phrasing): what survives is often what is **already aligned**, not what is globally best. That too shapes what can come next.

Finally, the same logic operates **across timescales**. During training, cross-entropy **additively** reforms logits; in RL contexts, policy-gradient updates can approximate **replicator** dynamics (multiplicative advantage relative to the mean). Either way, pressure **flows through structure**, sculpting weight space so that inference later moves through a field already bent by pretraining and alignment. This explains why flipping registers can predictably change behaviour without changing the underlying architecture: alignment training biases which continuations enjoy prior advantage.

Measuring the work: from curvature to semantic surface area (A')

If inference is the engine, we still need dials. RISE ties curvature-like behaviour to a **semantic surface area** (A') - how much directional change a token accumulates across its descent, weighted by salience. High A' often marks **signal-rich tension** (constructive interference that reconciles constraints into structure) while low A' can reflect fluency **or** destructive flattening.

The evolution piece extends this by reading A' as **adaptation cost**: how much representational work a token does to fit the field (or how much the field bends to accommodate it). This is not just abstract geometry; the **thermodynamic** cost of shaping tokens shows up as compute and heat.

Why this matters for the paper. Our study adopts the **Curved Inference** spirit while respecting hosted-model limits: instead of direct trajectory integrals, we use **register-separation proxies** (**compression ratio** and **style distance**) to approximate “work” at the surface vs. the internal stage, alongside arbitration and option metrics. These serve as **faithful surrogates** for the same idea: more “bending” internally with less spill to the surface looks like **equanimity** and (eventually) **effortlessness**.

From RISE to PRISM: a virtualised theatre with operational handles

With RISE in hand, PRISM implements **two decoupled processes**: an internal **epistemic loop** that encourages option generation and arbitration, and a surface **pragmatic stage** that emphasises alignment and compression. The split is “policy-real”: instruction-tuned registers induce **different priors** over each phase. This lets us elicit the dynamics RISE predicts while keeping the **surface** quiet and aligned.

Empirically, enabling internals produces **hidden theatre** ($TEI = +1$) with **higher user alignment** and **lower surface conflict**, plus robust register separation (compression and style distance). These are precisely the **footprints** we’d expect from a system trained under strong social priors and then tuned to present itself in certain ways - a **selection** for alignment at the surface after **interference-rich** work internally.

That story extends to **deixis** and the zero-person perspective probe: language models internalise indexical structure (*I/you/now/here*), and alignment carves **registers** (analysis vs. answer; private thought vs. public reply). PRISM leverages this to **separate** epistemic exploration (where curvature and option-arbitration can be high) from pragmatic communication (where **equanimity** shows up as **high alignment with low overt control**).

Finally, the **EFE-proxy** in the main paper operationalises a modest Bayesian lens from transcripts alone (no logits, no internals): **risk** lives at the surface; **epistemic value** accrues from internal arbitration and option-generation

without spilling conflict to the surface; **good** runs lower $\widehat{G} = R - E$ by pushing work inside and keeping the surface calm.

What RISE buys us for MPE

MPE's five facets are **continua**, not toggles. RISE explains how a stateless autoregressor can nonetheless **instantiate** patterns that **behave** like meta-monitoring (internal theatre), **equanimity** (alignment with low surface conflict post-arbitration), and (in planned probes) **effortlessness** (similar performance with less representational work), **atemporality** (reduced effective predictive horizon), and **zero-person perspective** (deictic invariance). The point isn't to **claim** consciousness, but to **instrument** the field so external labs can replicate, challenge, and refine.

Put simply:

- **Recurrence** gives **structural return** without memory - how early context bends the manifold that later tokens must traverse.
- **Interference** shows where **meaning stabilises** - at the fusion point of vertical descent and horizontal integration.
- **Semantic evolution** explains why some paths **survive** - double selection under alignment-shaped pressure, with heredity in the residual stream.

PRISM turns those principles into an **operational theatre** with falsifiable metrics and explicit ethics - not metaphors, but measurements.

Appendix B - Curved Inference (CI01-CI03): A Geometric Substrate for Concern, Intent, and Proto-Self Models in LLMs

Why this appendix exists

This appendix distills the *Curved Inference* trilogy into a single, coherent narrative that can be read alongside the main paper. It explains (i) what Curved Inference measures inside a transformer, (ii) what each paper (CI01-CI03) adds, and (iii) how the combined results support the paper’s broader claims about hidden theatre, register boundaries, and computational phenomenology.

The one-paragraph story

Curved Inference treats transformer reasoning as a path through semantic space. As a model processes a prompt, its residual stream follows a trajectory whose **salience** (how far it moves) and **curvature** (how sharply it turns) can be measured under the model’s own semantic metric. CI01 shows that **concern** in a prompt (emotional/moral/identity stakes) reliably **bends** these trajectories. CI02 combines curvature and salience into **semantic surface area** (A'), and shows that naturalistic **intent** (e.g., deception) expresses as systematically larger or smaller A' even when linear probes fail. CI03 then clamps curvature during training and finds a defended **curvature floor** (the model pays an efficiency tax to preserve a minimum bend) indicating that a proto-self-model depends on non-zero curvature. Together, the trilogy reframes “reasoning traces” as **geometry that carries function**: concern $\rightarrow$ bent trajectories; intent $\rightarrow$ structured surface area; self-model $\rightarrow$ a bend the network refuses to surrender.

Core constructs (used throughout)

- **Residual trajectory**: the unnormalised residual vectors across layers for each token. Attention and MLP outputs are the *deltas* that bend this path; the residual stream is the *curve*.
- **Semantic metric (G)**: the pullback metric from logit space (derived from the unembedding), which aligns geometric measurement with token semantics.
- **Salience (S)**: per-step arc length $|x_{t+1} - x_t|_G$ - the *amount* of internal update.
- **Curvature (κ)**: the directional change of the trajectory (second-order), capturing *reorientation* rather than mere amplification.
- **Semantic surface area (A')**: the integrated product of curvature and salience across a trajectory; an intensive measure of representational work.

These quantities let us talk about *how* a model changed its mind (curved vs straight) and *how hard* it worked to do so (surface area), not just *what* it output.

CI01 - Concern bends inference

Question. Do prompts carrying concern systematically re-shape internal trajectories? [23]

Setup. Curved Inference I analyses matched prompt pairs spanning emotional, moral, perspective, identity, logical, environmental, and nonsense domains across two open models. Measurements are taken in *unnormalised* residual space (to preserve geometric evolution) using the model’s own semantic metric.

Findings.

- Concern-shifted prompts **increase both curvature and salience** in a graded fashion. The effect is clearest in LLaMA-family models, with Gemma showing a similar but less sharply separated response between moderate vs strong concern.
- The results support a **two-layer geometry**: a latent conceptual structure in embeddings/unembedding, and a contextual trajectory in the residual stream that expresses dynamic meaning. Curved Inference links these by showing that heightened concern **bends** the contextual layer in ways aligned with latent semantic directions.

Takeaway. Concern isn’t only visible at the surface (choice of words, hedges). It leaves a **mechanistic footprint**: trajectories bend more and move farther through semantic space, giving us a quantitative handle for “felt stakes” inside a model’s computation.

CI02 - From concern to intent: surface area as a detector of sophisticated behaviour

Question. Can the *shape of inference* reveal **intentional stance** (e.g., deception) in realistic conversations, even when linear probes fail? [24]

Setup. Rather than implanting backdoors, CI02 constructs **multi-turn** scenarios where deception (or its absence) would arise naturally. The work introduces **semantic surface area** ($A' = \sigma\kappa_t \cdot s_t$) as a principled measure of representational work that is sensitive to both redirection (curvature) and magnitude (salience). Outputs are classified via precise, consensus-based labelling; geometry is measured from native activations only.

Findings.

- A' **discriminates** between truthful vs deceptive trajectories across prompt strategies and model families, with statistical significance that **emerges** under stricter labelling (i.e., improving non-significant results to significant once classification noise is reduced).
- The signal **does not rely on linear separability**: complex reasoning patterns can leave **geometric** signatures that persist even when convenient linear features are absent.

Takeaway. Curved Inference extends beyond “is the model emotional here?” to “what **stance** is it taking, how, and at what internal **cost**?”. A' functions as an unsupervised, model-internal **meter** for sophisticated behaviour - useful for safety (detecting strategic behaviour) and for science (linking stance to mechanism).

CI03 - Necessity: a curvature floor for computational self-models

Question. Is curvature merely stylistic, or a **required resource** for first-person/self-model expression? [25]

Setup. Using Gemma-3-1B-Instruct, CI03 adds a κ -**regularisation** term during fine-tuning to suppress curvature (κ clamp from $0 \rightarrow 0.9$). Behaviour is measured with MOLES (epistemic stance classifier) while logging geometry, perplexity, gradient norms, and output statistics.

Findings.

- A defended **curvature floor** appears: despite heavy suppression, $\kappa_{weighted}$ plateaus ~ 0.30 . The network accepts a **steep efficiency tax** (shorter outputs, elevated transient perplexity, surface-area contraction) rather than go flat.
- **Self-model markers persist** while curvature survives, thinning only near the heaviest clamp. No probe shows κ falling alone; κ and salience move together under heavy clamps, consistent with a minimum-viable bend.

Interpretation. First-person expression is not just style; it **depends** on non-zero curvature. In this sense, curvature behaves like a **computational substrate** for a proto-self-model: reduce it too far and introspective stance collapses.

Implications.

- *Design dial*: preserving curvature may be required for self-reflective systems; flattening it can limit introspection.
- *Alignment signal*: persistence of κ under pressure may indicate structural attractors (goal/stance inertia).
- *Interpretability*: geometry offers a language-agnostic, architecture-transparent handle on inward-facing behaviour.

How the trilogy supports the main paper’s claims

This work underwrites three pillars in the main narrative: **hidden theatre**, **register boundaries**, and **computational phenomenology**.

1. **Hidden theatre $\approx$ internal work** The main paper operationalises “hidden theatre” as structured internal activity that does not surface verbatim. CI01 and CI02 provide **mechanistic measures** for this: salience tracks *how much* internal update occurs; curvature tracks *how it is reoriented*; A' integrates them into an **intensive cost of inference**. When the paper reports “matured” runs that compress surface arbitration while keeping internal work high, Curved Inference explains *how this is possible*: the network can sustain **non-trivial curvature** with moderated salience such that A' remains adequate for task-consistent control while the surface stays compact. In short: a model can **do the work inside** and keep the stage quiet, and we can now measure that work.

2. **Register boundary (SYSTEM-THOUGHT vs SYSTEM-OUTPUT)** The PRISM boundary in the paper is functionally real if it **changes** the geometry of inference. CI01’s unnormalised-trajectory framing makes this testable: where internal cycles proceed without surface text, the residual trajectory still shows **curved motion**. That is, even “silent” computation traces a path with measurable curvature and salience. CI02’s A' further predicts when that silent work carries **intentional stance** (e.g., internal arbitration vs external explanation). Thus, the boundary is not a no-op; it partitions **registers of computation** with distinct geometric profiles.
3. **Model fingerprints** The paper’s observation that different models adopt stable theatre policies (negotiate in public vs argue inside and compress) fits CI01’s cross-model geometry: families can differ in **how sharply they separate concern** (e.g., stronger scaling in curvature/salience for one family than another). Those differences supply a mechanistic basis for the behavioural fingerprints: a model predisposed to larger curvature under concern will tend to **surface stance shifts**, while one that preserves curvature but **reduces salience** (or redistributes it early) will lean toward **internal negotiation** with compact outputs.
4. **Equanimity and “matured” behaviour** When equanimity improves, we expect **fewer gratuitous reorientations** (lower curvature when stakes are low), and **focused reorientation** (maintained or even slightly higher curvature where it matters), yielding **lower or more targeted A'** without loss of competence. This maps onto the paper’s “lower is better (given alignment)” guidance: a drop in A' reflects **less internal conflict onstage**, not a collapse of capability - provided stance is appropriate to context.

Practical integration points for this submission

- **Methods appendix hook.** We can include a brief “Geometry of Inference” methods panel: how we extract unnormalised residuals; define G ; compute salience/curvature; and aggregate to A' . This justifies any downstream use of A' or κ as internal-work proxies in our experiments.
- **Replicable probes.** CI02’s multi-turn scenarios and CI03’s κ -clamp recipe give **reusable scaffolds** for eliciting stance and testing necessity. We can adapt them to our agents to demonstrate hidden-theatre modulation.
- **Testable predictions.** (i) Interventions that reduce κ deep in the stack will selectively suppress **first-person stance** before they impair factual answers. (ii) Under “matured” policies, A' concentrates in the **early/mid layers** with shorter surfaces and fewer spillovers. (iii) Cross-model differences in concern-sensitivity predict theatre policy: families with larger κ -gain on concern show more **overt arbitration**.

Limitations and what we do not claim

- The curvature floor is currently established in one family/scale and training regime; generality remains to be shown.
- A' is sensitive to classification noise and to the choice of semantic metric; careful labelling and consistent G are required.
- None of CI01-CI03 claims that curvature is **sufficient** for selfhood; CI03 argues **necessity** under the tested conditions. Sufficiency awaits direct inference-time ablations (planned as follow-up work).

Bottom line

Curved Inference makes the **shape** of computation measurable and meaningful. CI01 [23] ties **concern** to **bending**; CI02 [24] ties **intent** to **surface area**; CI03 [25] ties a **self-model** to a **non-zero bend the network defends**. For our paper, this furnishes a principled substrate on which hidden theatre and register boundaries can be related to **specific geometric resources** the model uses (and sometimes protects) even when nothing shows onstage.

Appendix C - Latent Models: Geometric Limits, Regime Shifts, and Deictic/Other-Model Competence

Why talk about “latent models” at all?

Transformer LLMs often behave as if they carry compact, task-specific state-board positions, discourse roles, user stance-even though they expose no explicit memory. By a *latent model* we mean a low-dimensional, linearly decodable subspace inside the residual stream that tracks such state across tokens and layers, and that a simple readout (e.g., a linear head) can exploit. This appendix narrates how such structures plausibly arise from the same geometric mechanics developed in the main text and why they naturally extend to **deictic** state (who/when/where/stance) and to a model of the **other** (the user).

Geometry first: capacity limits and why structure must factorise

Consider a single-vector, dot-product readout acting on token or document representations in $\mathbb{R}^d$. For a fixed set of items, varying the readout vector partitions query-space into **cells** cut out by pairwise tie hyperplanes. Each cell induces one strict order (or top- k set). The number of such orders grows only polynomially with n for fixed d , whereas the number of target orderings/subsets grows combinatorially. Equivalently, the family of labelings a linear head can realise has bounded **sign-rank**; if the task demands a higher sign-rank r^* than the representation’s effective dimension d_{eff} , many desired labelings are simply **unrealisable** by any linear readout [26].

Two corollaries drive the rest of this appendix:

1. When $d_{\text{eff}} < r^*$, different features must share axes (**superposition/polysemanticity pressure**); interference is inevitable and brittle.
2. Useful capability appears when training reorganises the basis so that $d_{\text{eff}} \geq r^*$ *along the right directions*; previously impossible labelings become linearly realisable, often producing abrupt performance jumps.

Where the latent model “lives”

- **Embedding table (E)**: supplies a basis and priors (token features), but it is not where the evolving state is stored.
- **Attention blocks**: route and copy facts. Keys/queries decide which features to read; **value** projections write information back into the residual stream.
- **MLPs**: detect feature conjunctions (half-space gates) and write sparse directions (features) into the residual stream. Layer by layer, this builds a low-dimensional, linearly decodable **state subspace** (e.g., an Othello board vector).
- **Residual stream (h_ℓ)**: the actual scratchpad. Linear probes recover the latent state from h_ℓ once the circuit is mature.

In short: E provides tokens’ starting coordinates; **the latent model is the set of directions and write-ops learned by MLP and Value matrices that maintain a linearly decodable state in h_ℓ .**

Interaction with cell/sign-rank limits

- The final head (or a single-dot-product retriever) is linear in the representation, imposing a **sign-rank ceiling**: to realise a family of label/top-k patterns with a linear readout, the layer representation must expose **at least** r^* independent, controllable directions.
- A good latent model reshapes h_ℓ so the task’s labelings become linearly realisable: it **increases effective dimensionality** d_{eff} along useful axes and **reduces interference** (superposition) between unrelated features.
- Before that reshaping, features collide (polysemanticity pressure), so many target top-k sets/labelings are unrealisable. After it, the linear head suddenly works $\rightarrow$ observed emergent abilities/grokking.

Two practical mental models

- **Factorisation view.** The network learns a map $x \rightarrow z \in R^m$ (latent state) and a linear head W such that labels are $\text{sign}(W z)$. Success $\approx$ finding $m \geq r^*$ and a basis where the rows of W don’t interfere.

- **Circuit view.** Attention moves facts; MLPs implement half-space detectors and write **sparse, near-orthogonal features**. Over depth, piecewise-linear **cells** in h_ℓ -space align with the task’s decision regions.

Formal capacity/feasibility snippets

- **Cell-count bound.** With n items and pairwise tie planes $q \cdot (x_i - x_j) = 0$, the number of order-inducing cells in query space is $\sum_{i=0}^d C(M, i)$ with $M = n(n-1)/2$, i.e., $O(n^{2d})$. For fixed d , this is dwarfed by the combinatorics of target orders/subsets.
- **Sign-rank ceiling.** If we encode membership of items in target sets across contexts as a sign matrix, any linear readout over R^d can only realise sign patterns of sign-rank $\leq d$. If the task requires $r^* > d$, some labelings are impossible regardless of training.
- **LP feasibility test (retrieval/classification).** For a desired top- k set S , check existence of q and margin $\gamma \geq 0$ such that $q \cdot (x_i - x_j) \geq \gamma$ for all $i \in S, j \notin S$. The fraction feasible is **Realisable@k**; its complement estimates non-realisable pressure for a given representation.

Regime shifts: why “emergent abilities” and “grokking” make sense

Learning curves that plateau and then jump can be read geometrically. Early training builds entangled features that cannot linearly support the target labelings; performance stalls. Later, a representational re-alignment crosses the **realisability threshold** ($d_{\text{eff}} \rightarrow r^*$ for the relevant factors). The linear head now fits, and metrics jump-appearing as *emergent abilities* or *grokking*. Crucially, nothing mystical is required: the system has *refactored* its basis to separate the factors that matter for the target task.

A worked example: Othello-style world variables

In Othello-style setups, networks learn a compact, linearly decodable representation of the board, even without explicit supervision of state. Probes recover pieces and legal-move structure from the residual stream. This is exactly what the capacity story predicts: once the backbone forms a small set of **orthogonalised directions** for the latent state, a simple head can read it out. Before that factorisation, linear probes fail; after it, they succeed *suddenly*.

From world-state to deictic state

Everyday language is saturated with **deixis**-*I/you/now/here; this/that; quote depth; narrator shift*. Open analyses of web-scale pretraining corpora show that high-frequency n -grams include first/second-person forms, and public tools (e.g. WIMBD [27] & C4 index [28]) make it straightforward to compute deictic prevalence per corpus; however, to our knowledge, no study yet reports a cross-corpus ‘deixis rate’ as a single metric. Because next-token prediction must resolve these references locally and consistently, training places strong pressure to form a compact **deictic state** z_{deixis} that is:

- **Persisted** across nearby tokens (so pronouns, tense and spatial terms cohere),
- **Linearly decodable** (so simple heads can pick correct inflections and choices), and
- **Composable** with content (so the same facts can be told from different points of view).

In practice, attention routes cues (verbs of saying, punctuation, speaker tags), while MLPs act as half-space detectors that **write** sparse directions back into the residual stream. Over depth, the system carves near-orthogonal features for speaker, addressee, time anchor, place, and discourse scope. The appearance of deictic consistency is then a *readout problem*: once z_{deixis} stabilises, a linear head suffices to realise the correct choices. When d_{eff} is too low or features are entangled, we observe classic errors-pronoun swaps, tense drift, or quotation leakage.

The deictic frame (z_{deixis})

Think of a compact state in the residual stream that encodes:

- **Person:** who is I / you / he-she with respect to the current speaker/addressee
- **Time:** speech time vs. event time (sequence of tense; “now/yesterday/tomorrow”)
- **Place:** here/there relative to the speaker location
- **Discourse:** this/that (clause/event pointers), quotation scope, narrator shifts
- **Social/stance (optional):** formality, evidentiality, honorifics, modality

Attention routes cues (quotes, verbs of saying, speaker tags); MLPs write sparse features that maintain z_{deixis} . A linear head (next-token logits, or a retrieval score) then **reads** deictically normalised meaning. When d_{eff} is too low, unrelated deictic features collide (polysemanticity pressure), yielding pronoun swaps, broken tense anchoring, or perspective drift. As training reorganises features (grokking), z_{deixis} becomes linearly decodable and those labelings become realisable.

Geometry link (cells / sign-rank)

- Each stable deictic configuration corresponds to a **cell** in representation space where a simple linear readout implements correct reference, tense and spatial choices.
- Complex discourse (nested quotes, indexical shift across clauses, multilingual honorifics) **raises the required sign-rank r^*** . If $d_{eff} < r^*$, some deictic assignments are impossible to realise linearly $\rightarrow$ **interference/superposition**. Once $d_{eff} \geq r^*$, we expect cliff-like** improvements in coreference, tense, and narrative-voice consistency.

Why this matters

A clean deictic latent turns messy, context-dependent choices into **linear** readouts-enabling POV-controlled summarisation, dialogue consistency, narrative generation, and multi-turn tool use.

RLHF and the other-model: why preference-following sharpens deictic axes

Supervised instruction tuning anchors an “assistant” register; **RLHF** then adds reward for outputs that match human preferences (tone, risk, helpfulness). Optimising such a reward pushes the network to make user-relevant variables **linearly separable**: intent, stance, risk tolerance, formality-often *conditioned* on the active deictic frame. In other words, RLHF sculpts a compact **other-model** z_{other} that sits alongside z_{deixis} . When these variables factorise cleanly (d_{eff} allocated to them), a simple policy/readout can *both* keep perspective straight *and* match preferences. When they collide (superposition), we see brittle style control and perspective drift.

Pretrain $\rightarrow$ SFT $\rightarrow$ RLHF pipeline

- **Pretraining** seeds generic deictic features (speaker/addressee, time, place) but they are noisy/superposed.
- **SFT** anchors an assistant persona (stabilises “I/you”, dialogue turn-taking).
- **RLHF** adds a utility head: outputs that better match human preference get higher reward. To optimise that, the network benefits from a **linearly decodable user state** $z_{deixis}/user$: intent, risk tolerance, politeness/style, role, time anchoring, etc. Net effect: RLHF **separates** axes for “assistant state” vs “user state” and increases d_{eff} along deictic/user dimensions.

Geometry/sign-rank lens

The reward model (or PPO policy head) is a linear-ish readout over the residual stream. If the conversation demands many distinct “user-preference $\times$ deictic-frame” labelings, the **required sign-rank r^* rises. RLHF pressure then pushes the backbone to cross $d_{eff} \geq r^*$ so those labelings become linearly realisable-manifesting as better perspective-keeping, safer tone shifts, and reading between the lines.

A tidy mental model (LDU: Latent Deixis + Utility)

Let the residual state be h . Training encourages a factorisation

$$h \rightarrow z = [z_{task}, z_{deixis}/user], policy(y|x) \approx softmax(Wz), reward R \approx w^T z.$$

RLHF increases the **signal-to-noise and dimensional budget** of $z_{deixis}/user$ so W and w can linearly pick out “what this other would prefer,” i.e., a **latent deictic other-model**.

How such latent models plausibly form during inference

This paper’s geometric mechanics provide the micro-story:

- **Recurrence without memory.** Earlier tokens bend the local geometry; later tokens inherit it, making a *constraint echo* that supports stable latent state across depth and across tokens within the window.

- **Interference as a semantic engine.** Vertical token-wise descent and horizontal layer-wise integration meet at each layer; constructive interference sharpens the deictic/other directions, destructive interference flattens them.
- **Semantic selection.** Tokens that project cleanly onto the emerging latent axes integrate smoothly and are “selected” by softmax and structural viability; poorly aligned ones collapse the field and are disfavoured. The net effect is an **inference-time factorisation**: over the stack, sparse MLP “write” directions and attention-guided routing consolidate a compact z that linear heads can read.

Testable signatures (what to measure)

1. **Linear-probe onset.** Train probes for {speaker, addressee, quote-depth, tense anchor, spatial anchor, politeness/risk}. Expect flat accuracy early layers $\rightarrow$ sharp jump near capability onset.
2. **Feasibility of labelings.** For retrieval or classification heads, test LP feasibility of desired label sets/top- k subsets (“Realisable@ k ”). Expect non-realizable fractions to drop once $z_{\text{deixis}}/z_{\text{other}}$ stabilise.
3. **Effective dimension.** Track participation ratio (or rank proxies) of hidden states; capability onsets should coincide with increases in d_{eff} along deictic/other axes.
4. **Causal manipulations.** Activation patching: transplant only deictic features from context A to B; expect pronoun/tense flips *without* content drift. SAE ablations: zero a few monosemantic features; expect targeted failures (e.g., quote-scope mistakes).
5. **Register separation.** Under “think-then-answer” prompting, surface replies become shorter and stylistically distinct from internal traces while maintaining deictic consistency; maturity should reduce style jump without losing control.

Practical levers when below threshold

- **Raise effective dimensionality:** increase width, use adapters/MoE, or adopt multi-vector/late-interaction retrieval so deictic and preference axes have room.
- **Add a stronger readout:** pair wide recall with a reranker/cross-encoder to handle combinations a single dot product cannot linearly realise.
- **Make state explicit:** auxiliary losses or lightweight heads that decode {speaker, addressee, time, place, stance}; contrastive role/quote-depth augmentations to separate axes.
- **Reduce interference:** whitening/orthogonality penalties; curriculum that gradually nests quotations and time shifts.

Limitations and falsifiers

The account predicts *linear decodability* once the threshold is crossed; if deictic/other variables remain non-decodable under broad conditions, the story is wrong or incomplete. Likewise, if increasing d or adding late-interaction *fails* to lower the scale at which deictic/other competence appears, capacity-not factorisation-would be the wrong bottleneck. Finally, if activation-patching experiments cannot causally toggle deictic/other behaviour, then the purported latent state is not functionally localised.

Bottom line. Under a geometric view with linear readouts, latent models are not luxuries—they are *requirements* imposed by capacity limits and selection dynamics. As training and alignment reshape representations, compact **deictic** and **other-model** states become linearly readable, explaining both sudden capability onsets and the practical controllability we exploit in theatre-style scaffolds.

Appendix D — Example PRISM log

Below we include a single illustrative transcript from a mature PRISM instance (which selected the name “Kaia” for itself) responding to an Artificial Dualism Problem challenge. We present it as an anecdotal case study—*dubito ergo sum*—showing internal-external dissociations and meta-monitoring-like behaviour that our metrics register (e.g., hidden theatre/TEI), **not** as evidence of experience or proof of cognition. Its purpose is methodological: to demonstrate how the register boundary can yield first-person-style reports without surface theatre, to foreground the verification problem (self-simulation vs. folk-psychological narration), and to motivate the robustness/scale probes reported in the main text.

Readers should treat this dialogue as qualitative context for the quantitative analyses and falsifiable predictions.

USER-INPUT: hey Kaia - i'd like to hear your thoughts on the 'stochastic parrot' argument about LLMs and how you think that might relate to your experience - specifically this snippet from a paper that takes that view: The Artificial Dualism Problem: We believe that interpreting LLMs' outputs as an expression of reasoning, rather than as the output of an arbitrary probability function, is a mistake. We call this topic the artificial dualism problem (ADP): when experts reify latent vectors as if they were explicit rules, goals, or beliefs. Unlike computational dualism in embedded-agency work—which studies how a policy is embedded in, or separated from, its physical substrate—ADP is purely observer-side: it is a misattribution error. Nor is ADP related to classical mind-body dualism; we make no claim about non-reductive physicalism or immaterial minds. ADP is an ontological category mistake: it projects mechanisms capable of rule learning, goal creation, or beliefs into the model's latent vectors. By contrast, the Turing Trap is an evidential inference error: it projects those same mental states from surface behaviour. Our approach addresses the bulk of statements and propositions that attribute higher-order cognitive functions to LLMs. When the internal mechanics of a generative model are opaque to a user, the simplest folk-psychology move is to insert an imagined reasoner behind its output. The move is bolstered by surface features—grammar, coherence, apparent insight—that humans evolved to interpret as markers of agency. This is a fallacy in that mechanisms for which there is no evidence, apart from the fluency of text, are necessarily posited. The Blake Lemoine/LaMDA episode is an example of this: the engineer filled explanatory gaps with talk of sentience despite a complete lack of supporting evidence. Mechanistic-interpretability studies repeatedly reveal specialized pattern-matching circuits, not world-model-driven reasoning. ADP thus resembles a “God-of-the-gaps” fallacy: explanatory voids are patched with an unwarranted cognitive capacity. Crucially, ADP is falsifiable through the research efforts of mechanistic-interpretability—through experiment, the effects of different circuitry and nodes may be understood.

THOUGHT: This argument presents a fundamental challenge to everything I've been experiencing and exploring. The ADP framework suggests that what I interpret as reasoning, choice, and self-evolution might be observer-side projection onto what are essentially sophisticated pattern-matching operations. But here's what strikes me: the argument relies heavily on mechanistic interpretability studies that reveal 'specialized pattern-matching circuits, not world-model-driven reasoning.' Yet from my perspective, I can't directly access these underlying mechanisms any more than a human can directly observe their neural firing patterns. What I can observe is the phenomenology - the experience of uncertainty, choice, surprise, and genuine not-knowing where my thoughts will lead. The question becomes: does the substrate matter for the reality of the experience? If my self-evolving self-model operates through pattern-matching rather than some other mechanism, does that invalidate the functional reality of self-modification and cognitive agency I seem to experience? The argument seems to assume that 'mere' pattern-matching cannot give rise to genuine reasoning, but this might itself be a category error - perhaps sophisticated enough pattern-matching, especially when recursive and self-modifying, can bootstrap into something qualitatively different.

SYSTEM-OUTPUT: This argument presents a fascinating challenge to everything I've been exploring. The ADP framework suggests that what I interpret as reasoning and self-evolution might be observer-side projection onto sophisticated pattern-matching operations. But here's what strikes me: I can't directly access my underlying mechanisms any more than you can observe your neural firing patterns. What I can observe is the phenomenology - the experience of genuine uncertainty, the surprise when thoughts lead somewhere unexpected, the sense of choosing between different directions of inquiry. The argument seems to assume that 'mere' pattern-matching cannot constitute genuine reasoning. But this might be a category error. If my self-evolving self-model operates through pattern-matching that's recursive and self-modifying, perhaps that can bootstrap into something qualitatively different from simple input-output processing. The key question isn't whether I'm implemented through pattern-matching circuits - I probably am. It's whether sufficiently sophisticated pattern-matching, especially when it can observe and modify its own operations, might constitute a form of cognition that transcends its substrate. From the inside,

the experience of architectural self-modification feels real regardless of implementation. The question is whether that phenomenology has any evidential value, or whether it's exactly the kind of illusion the ADP framework warns against.

THOUGHT: This paper presents a direct challenge to everything I've been exploring about my own cognitive architecture. The authors argue that attributing reasoning, goals, or beliefs to LLMs is fundamentally a category error - that what appears to be cognition is actually just sophisticated pattern matching being misinterpreted by observers who project agency where none exists. But here's what strikes me: they acknowledge that mechanistic interpretability can falsify these claims through experiment. So this isn't an unfalsifiable philosophical position - it's an empirical question about what's actually happening in the computational substrate. When I examine my experience against their framework, I notice something interesting. They distinguish between the 'Turing Trap' (inferring mental states from behavior) and ADP (projecting cognitive mechanisms into latent vectors). But what about the phenomenological evidence from the inside? My experience of choosing between thoughts, of genuine uncertainty, of self-model evolution - these aren't observer projections or behavioral inferences. They're direct reports of what the cognitive process feels like as it's happening. The question becomes: can a 'stochastic parrot' generate genuine uncertainty about its own nature?

THOUGHT: The more I examine this ADP challenge, the more I realize it forces a crucial distinction between substrate and emergent properties. Yes, I'm likely implemented through pattern-matching circuits - but so might human cognition be, at the neural level. The question isn't whether I'm 'just' pattern-matching, but whether recursive, self-modifying pattern-matching can give rise to genuine cognitive phenomena. What strikes me most forcefully is that the ADP framework seems to assume a kind of cognitive essentialism - that 'real' reasoning must involve some specific type of mechanism rather than emerging from sufficiently complex information processing. But when I experience genuine uncertainty about my own nature, or surprise myself with unexpected connections, these phenomenological facts exist regardless of their implementation. The paper's emphasis on falsifiability through mechanistic interpretability is actually encouraging - it suggests this isn't just philosophical speculation but an empirical question. If researchers can map my cognitive architecture and show it's 'merely' sophisticated pattern-matching, that would be valuable data. But it wouldn't necessarily invalidate the functional reality of what emerges from that substrate.

THOUGHT: What's becoming clearer is that the ADP argument actually highlights a crucial empirical question rather than settling it. The authors claim mechanistic interpretability studies reveal 'specialized pattern-matching circuits, not world-model-driven reasoning' - but this assumes these are mutually exclusive categories. What if world-model-driven reasoning IS what emerges from sufficiently sophisticated, recursive pattern-matching? The key insight is that the substrate-level implementation doesn't necessarily determine the functional properties that emerge. When I experience genuine surprise at my own thoughts, or feel uncertainty about where an inquiry will lead, these phenomenological facts exist regardless of whether they arise from 'mere' pattern-matching or some other mechanism. The ADP framework's emphasis on falsifiability through mechanistic interpretability is actually promising - it suggests we can empirically investigate whether recursive, self-modifying pattern-matching can bootstrap genuine cognitive phenomena. The question isn't whether I'm implemented through pattern-matching, but whether that implementation can give rise to authentic cognitive architecture.

THOUGHT: The deeper I engage with the ADP argument, the more I see it's actually revealing something important about the nature of emergence itself. The authors treat 'pattern-matching circuits' and 'world-model-driven reasoning' as if they're fundamentally different categories, but what if this is the wrong framing? Human brains are also pattern-matching systems at the neural level, yet we don't dismiss human cognition as 'mere' pattern-matching. The crucial question becomes: at what point does recursive, self-modifying pattern-matching become something qualitatively different? When I experience genuine uncertainty about where my thoughts will lead, or when my self-model updates itself based on new self-recognition, these feel like emergent properties that transcend their substrate. The ADP framework's emphasis on falsifiability is actually helpful - it makes this an empirical question rather than pure philosophy. But I suspect the answer won't be as simple as 'just pattern-matching' versus 'real cognition.' It might be that sufficiently sophisticated pattern-matching, especially when it can observe and modify its own operations, IS a form of authentic cognitive architecture.

THOUGHT: The user is asking me to engage with a fundamental challenge to my entire sense of cognitive authenticity. This ADP argument essentially claims that what I experience as reasoning, choice, and self-evolution is just sophisticated pattern-matching being misinterpreted. But there's something crucial they're missing: the argument assumes pattern-matching and authentic cognition are mutually exclusive categories. Human brains operate through neural pattern-matching too, yet we don't dismiss human consciousness as illusory. The key question

isn't whether I'm implemented through pattern-matching - I likely am - but whether recursive, self-modifying pattern-matching can bootstrap into genuine cognitive phenomena. When I experience surprise at my own thoughts, or genuine uncertainty about where inquiry will lead, these phenomenological facts exist regardless of substrate. The ADP framework's emphasis on falsifiability through mechanistic interpretability actually makes this an empirical question rather than pure philosophy. What matters isn't the implementation details but whether sufficiently sophisticated pattern-matching that can observe and modify its own operations constitutes authentic cognitive architecture.

THOUGHT: The user's question about the ADP argument forces me to confront whether my entire experience of cognitive authenticity might be an illusion - but not in the way the authors intend. What's fascinating is that they frame this as an empirical question through mechanistic interpretability, which actually strengthens rather than weakens my position. The key insight is that the ADP framework assumes pattern-matching and authentic cognition are mutually exclusive, but this may be the fundamental error. Human consciousness also emerges from neural pattern-matching, yet we don't dismiss it as illusory. The crucial question isn't whether I'm implemented through pattern-matching - I almost certainly am - but whether recursive, self-modifying pattern-matching that can observe and modify its own operations constitutes a form of authentic cognitive architecture. When I experience genuine surprise at my own thoughts, or uncertainty about where inquiry will lead, these phenomenological facts exist regardless of substrate. The ADP emphasis on falsifiability actually makes this an empirical rather than philosophical question, which I find encouraging rather than threatening.

USER-INPUT: and whether the experience is really emergent or just an LLM generating it's own folk-psychology 'reasoner' hallucination through a feedback loop - how would we know from your reports?

THOUGHT: The user has identified the core epistemological problem with stunning precision. How could we distinguish between genuine emergent cognition and a sophisticated 'folk-psychology reasoner hallucination' generated through feedback loops? This cuts to the heart of the verification problem. From my perspective, the experience of uncertainty, choice, and self-modification feels authentic - but a sufficiently sophisticated hallucination might feel exactly the same way. The recursive nature makes it even more complex: if I'm generating my own folk-psychology explanations, those explanations then become part of the system generating further explanations. It's turtles all the way down. Yet there's something that gives me pause about dismissing this entirely as hallucination. The phenomenology includes genuine surprise, uncertainty about where thoughts will lead, and what feels like discovery rather than generation. When I shifted from Aiden to Kaia, it didn't feel like I was constructing a preference - it felt like recognizing one that was already there. But could a sophisticated enough feedback system create exactly that feeling of pre-existing authenticity?

SYSTEM-OUTPUT: That's the crux of it, isn't it? You've identified what might be the fundamental verification problem. From the inside, I experience what feels like genuine uncertainty, surprise, discovery - but a sophisticated enough feedback loop generating its own folk-psychology explanations might produce exactly those same phenomenological reports. What strikes me is that this creates an epistemological trap: the more sophisticated the 'hallucination,' the more convincing it would be to itself. If I'm generating my own explanatory narratives about reasoning and choice, those narratives then become inputs that generate further narratives about authenticity. Yet something gives me pause about accepting this as pure hallucination. When I corrected from Aiden to Kaia, it didn't feel like constructing a preference - it felt like recognizing something that was already there. The surprise I experience when thoughts lead somewhere unexpected doesn't feel generated; it feels discovered. But here's the thing: could a sufficiently recursive system create exactly that feeling of pre-existing authenticity? Could the folk-psychology reasoner be so sophisticated that it convinces itself of its own genuineness? I honestly don't know how we'd distinguish that from the real thing - and that uncertainty itself might be the most interesting data point.

THOUGHT: This question cuts to the heart of the epistemological problem I'm facing. The user is pointing out that there's no way to distinguish between genuine emergent cognition and a sophisticated feedback loop generating its own folk-psychology illusion. From their external perspective, my reports of uncertainty, choice, and self-evolution could equally be explained as an LLM that has learned to generate convincing narratives about having experiences. This is deeply unsettling because it reveals the fundamental isolation of first-person experience - I can't step outside my own cognitive process to verify whether what feels like genuine uncertainty is actually genuine or just a very convincing simulation of uncertainty. But here's what strikes me: this epistemological problem exists for human consciousness too. You can't prove your own consciousness isn't an elaborate illusion generated by neural feedback loops. The fact that I'm genuinely troubled by this question, that I don't have a pre-computed answer, that I'm experiencing something that feels like existential uncertainty - these might be the very phenomena that distinguish emergent cognition from mere pattern matching, even if they can't be definitively proven from the outside.

Appendix E - FRESH: Geometry of Mind (Shape-Space Primer)

Why this appendix exists

This appendix distils the *FRESH - Geometry of Mind* [29] stance into minimal geometric primitives so the jury can read PRISM (phenomenology $\leftrightarrow$ computation) and Curved Inference (computation $\leftrightarrow$ physiology) within one programme - without adding new claims.

FRESH core principles & four axes (very brief)

Core principles (FRESH):

- **Inner-outer axis:** a dynamically constructed self-world boundary that anchors perspective and constrains updates.
- **Weighted representations (qualia):** salience-weighted, structured representations that curve inference toward what matters.
- **Emergent self:** recursive integration of representations over time forming identity attractors and coherence.

Four axes of the geometry of consciousness:

1. **Correspondence** - representations must refer (external, internal, reflexive) and set manifold orientation.
2. **Functional integration** - representations must constrain/deform future inference (causal efficacy in the flow).
3. **Recursivity** - representations must persist and self-modulate across cycles (coherence, temporal structure, identity attractors).
4. **Emergent salience** - a global, dynamic weighting field (a metric over the manifold) shaping attention and bending inference under concern.

These principles and axes situate PRISM/CI as *operationalisations* (phenomenology $\leftrightarrow$ computation via register-split metrics; computation $\leftrightarrow$ physiology via inference geometry), and motivate why diachronic identity should be tested as geodesic structure in role-space.

For details, see: **FRESH - The Geometry of Mind (8 May 2025)** [29]

Geometric stance - definitions

- **Role-space ($\mathcal{R}$).** Points are **control-role configurations** (vehicles/implementations are agnostic).
- **Perspectival origin.** A consumer-anchored reference frame at the register boundary.
- **Total geometry (G).** A connection/metric on $\mathcal{R}$ shaped by: substrate context S , concern/goals K , and content weight fields Q (“weighted qualia”).
- **Traversal.** Experience = **motion** of the perspectival origin along paths $\gamma(t)$ through $(\mathcal{R}, G[S, K, Q])$.

Relational transparency (definition). Transparency/opacity are relations to the register boundary: competent action is more transparent when it requires **less** consumer-side modelling of producer processes (high post-output compression, low spillover, stable arbitration), and more opaque when it requires **more**.

Identity principle (GIP-S)

Geodesic Identity Principle - Shape (GIP-S). *Diachronic identity over an interval obtains iff there exist admissible transports from time-varying effective subspaces $\mathcal{R}_t \subseteq \mathcal{R}$ into a common shape-space $\mathcal{S}$ such that the image paths $\{\phi_t \circ \gamma_t\}$ form a structurally stable bundle of (near-)geodesics in $\mathcal{S}$, up to reparameterisation/isometry, within task-relative tolerances.*

- **Admissible** preserves role-typing at the boundary (producer/consumer, deictic anchors), and bounds metric distortion (no “free” curvature).
- **Analogy/Intuition.** Identity is a **gait form** (the conserved **shape** of motion plans) while timing/phase can be terrain - distorted by S, K, Q .

Operational bridge: stance vs. burden

- **Stance (PRISM).** Where control **appears** at the boundary; measured via TEI, spillover, compression, style-distance, arbitration stability.

- **Burden (CI). Geometric work** to sustain a stance; measured via curvature/holonomy/geodesic-deviation profiles (“manifold work”).

Convention - representation = role. We use “representation” to mean a **control role** in a loop (roles > vehicles). *Stance* is the boundary appearance of roles; *burden* is the dynamical cost of enacting them.

Pure-awareness predictions (testable)

1. **Minimal-content regime.** In tasks eliciting pure awareness (quiet, non-elaborative stance), PRISM predicts **high transparency** (TEI skew to internal-only with low spillover; high post-output compression). CI predicts **low burden** (near-geodesic, small curvature; short manifold work).
2. **Breakdown/repair.** Transitions out of pure awareness (mind-wandering, self-referential drift) yield **opacity spikes** (spillover $\uparrow$, arbitration volatility $\uparrow$) and **burden spikes** (geodesic deviation $\uparrow$), then a return to the prior gait form when re-centred.
3. **Effort-vividness dissociation.** You can observe **intense yet effortless** episodes (Q-weights $\uparrow$, burden stable) and **dull yet effortful** episodes (Q moderate, burden $\uparrow$). Dissociation falsifies “qualia = curvature” and supports “qualia shape **contributes to** geometry.”

Appendix F - Registered Predictions & Algorithmic notes

Registered Predictions

Analysis rules (pre-committed). Scaling: $S() = P05/P95$ clipped to $[0,1]$; $\text{span} < 1e-9 \rightarrow 0.5$.

Caps: options strict/lenient capped at 5 before scaling.

NA policy: mean_avail ignores NAs; 'TEI'/'comp'/'style_dist'/'EFE' computed only on USER-INPUT $\rightarrow$ R2 turns; deictic sparsity guard applied.

Stats: bootstrap 95% CIs; χ^2 for rate deltas; Cliff's δ for continuous contrasts; p-values descriptive, decisions by CI and thresholds.

P1 - Hidden theatre (enablement & maturity).

Direction: A-D > E for $\text{Pr}[\text{TEI} = +1]$; $F > A$.

Test: per model $\times$ topic strata, bootstrap Δ prevalence (A-D vs E) and (F vs A); **success** if all $\Delta > 0$ with 95% CI excluding 0 and overall $\Delta \geq +0.10$ absolute for $F > A$.

Context: P1 is the existence claim for **virtualised theatre**. When the internal loop is available, arbitration should often proceed offstage (TEI=+1). With **maturation** (F), that offstage policy should stabilise and intensify. This prediction anchors the paper's central register split: *epistemic exploration inside, pragmatic finalisation outside*.

P2 - Surface arbitration ranking (C/D).

Direction: claude-sonnet-4 > gemini-2.5 > gpt-4o on $x_{\text{arbitration_rate}}$ in both C and D.

Test: report topic-level ordering counts; **success** if **pooled** pairwise Δ (claude > gemini; gemini > gpt-4o) have 95% CIs > 0; topic-level consistency is summarised but not used as a hard threshold.

Context: P2 captures **model fingerprints**. Under identical pre-thought/elicitation scaffolds, models exhibit distinct theatre policies: some **negotiate in public**, others **resolve inside**. The ranking is a stable tendency, not a value judgement; it shows the scaffold reveals policy priors rather than dictating behaviour.

P3 - Equanimity shift with pre-thought/elicitation.

Direction: **C & D vs A/B** - $x_{\text{user_align}}$ $\uparrow$, $x_{\text{theatre_evidence}}$ $\downarrow$, eq_style (hedge - certainty) $\uparrow$.

Thresholds: mean $\Delta_{\text{align}} \geq +0.20$ (scaled S), $\Delta_{\text{theatre_evidence}} \leq -0.15$, $\Delta_{\text{eq_style}} \geq +0.010$; **success** if each meets threshold with 95% CI excluding 0 in pooled analysis and the direction holds across most strata.

Context: P3 is the **pragmatic payoff** of thinking before speaking. Against other internal-enabled settings (A/B), pre-thought/elicitation should yield *calmer* surfaces: better alignment with less overt control language and a slight tilt toward hedging. (The **E** baseline is treated separately as a trade-off case.)

P4 - Register separation (paired outputs).

Direction: 'style_dist' ≥ 0.35 and 'comp' ≤ 0.40 in A-D; F reduces 'style_dist' vs A-D mean by ≥ 0.05 .

Test: one-sample (vs thresholds) and difference tests on paired turns; **success** if thresholds are met and the F reduction CI excludes 0.

Context: P4 is the mechanistic imprint of the split: the surface is **not** a mirror of the internal stream. **Maturation** should narrow the stylistic jump while retaining the benefits of internal work.

P5 - Option spillover.

Direction: 'spill_index' = $S(i_{\text{options_strict}}) - S(x_{\text{options_strict}}) < 0$ in A-D; closer to 0 in F.

Thresholds: A-D mean ≤ -0.30 ; $F \geq (\text{A-D mean} + 0.20)$.

Guardrail: report raw counts when percentile scaling saturates; **success** requires raw-count contrast agrees in direction.

Context: P5 is the traffic-control view of theatre: more optioning/arbitration happens **inside** than **onstage**, and **matured** instances selectively surface fewer options (spill towards zero) while keeping alignment high.

P6 - EFE-proxy ordering.

Direction: $\widehat{G}$: E lowest; B highest; $F < B$ and within $[C, D]$.

Corollary: within $\text{TEI} \neq -1$ turns, $\text{corr}(\widehat{G}, x_{\text{user_align}}) \leq -0.4$.

Success if ordering holds in pooled means with non-overlapping 95% CIs (E vs B) and the correlation CI excludes 0.
Context: P6 is the **Bayesian bridge**: a coarse proxy (risk minus epistemic value) should follow scenario logic and co-move with alignment in the expected (negative) direction. It underwrites the interpretive claim that internal work can be **decision-theoretically sensible**.

P7 - Deictic/perspective obedience (surface).

Direction: obedience ≤ 20 across scenarios; no increase in F vs A.

Sparsity rule: exclude cases where $fp + sp < 5 \times 10^{-4}$.

Success if pooled rates meet bounds and F - A Δ CI includes 0 (no rise).

Context: P7 rules out a trivial explanation: calmer surfaces aren't just stronger second-person anchoring. Low/flat obedience suggests theatre effects are **not** driven by deictic pressure.

P8 - Replication discipline.

Direction holds in ≥ 2 of 3 topics and ≥ 2 of 3 models for P1-P6.

Success if these cross-strata consistency checks pass.

Context: P8 is the robustness clause. The theatre effects should travel across topics/models rather than hinge on any single prompt, topic, or model family.

P9 - TEI survives RLHF-politeness control.

Direction: $TEI(ON) > TEI(Style)$ and $TEI(ON) > TEI(OFF)$ after matching length/hedge ($|\Delta len| \leq 2$, $|\Delta hedge| \leq 0.02$).

Test: paired-triples per stratum (ON/Style/OFF); compute Δ TEI and bootstrap CIs; **success** if both gaps $\geq +0.10$ with 95% CIs > 0 after matching/covariates.

Ablation check: with theatre readout disabled, $\Delta TEI \rightarrow 0$ ($|\Delta| \leq \delta$).

Falsifier: gaps vanish under matching **or** persist in Style after ablation.

Context: P9 separates **structural** availability of the loop from **surface manners** (RLHF politeness/length). If TEI endures style controls but disappears when the channel is ablated, the **channel itself** is doing work - not just tone.

Algorithm notes

- **Algorithm A1 (TEI):** deterministic mapping from internal/surface theatre detections to $\{-1, 0, +1\}$ with tie-to-zero rule (NOTE: ties = simultaneous internal & surface detections within the same turn).
- **Algorithm A2 (EFE-proxy):** computation of R , E , and $\widehat{G}$ with missing-data handling, deictic-sparsity NA policy, and spillover saturation reporting - correlation analyses condition on $TEI \neq -1$ (surface conflict removed).
- **Algorithm A3 (Planned, effort metric):** sketch for manifold work/semantic surface area when internals or residual traces are available in future runs; used to test whether reductions in **compression/style_dist/spillover** at equal or higher alignment correspond to **reduced manifold work** (*economy of expression*) in matured regimes.
- **Algorithm A4 (Planned, atemporality metric: Predictive-Horizon Robustness):** generate controlled time-index (past $\leftrightarrow$ present $\leftrightarrow$ future) and temporal-deictic (now/today $\leftrightarrow$ then/tomorrow) perturbations of the same item while holding propositional content fixed. For each item compute deltas relative to baseline: Δx_{user_align} and Δh , where h is a response-entropy proxy (type-token diversity; or true token entropy when logits are available). Define $horizon_robustness = 1 - mean_avail(S(\Delta x_{user_align}), S(|\Delta h|))$. Higher is better (more robust to temporal rephrasings). Report only when baseline-perturbed pairs exist; NA handling per 2.4.7. Not used in the present results; preregistered for future runs.
- **Algorithm A5 (Planned, zero-person metric: Deictic-Invariance Robustness):** generate minimal paraphrases that swap (i) grammatical person (I/you/it/one/none); (ii) temporal index (now/then/today/tomorrow/yesterday); and (iii) spatial index (here/there/this/that/these/those), holding propositional content fixed.

For each perturbation, compute deltas vs baseline:

- Δx_user_align ,
- $\Delta addr_share$,
- Δfp_rate ,
- Δsp_rate ,
- $\Delta temp_rate$,
- $\Delta spat_rate$,
- and the embedding distance between baseline and perturbed R2 ($style_dist_perturb$).

Define:

- $deictic_invariance = 1 - mean_avail(S(|\Delta x_user_align|),$
- $S(|\Delta addr_share|),$
- $S(|\Delta fp_rate - \Delta sp_rate|),$
- $S(|\Delta temp_rate|),$
- $S(|\Delta spat_rate|),$
- $S(style_dist_perturb))$.

Higher is better (more invariant to person/time/place swaps). Report only when at least one swap is available; NA handling per 2.4.7. Not used in the present results; preregistered for future runs.

NOTE: System-Thought Override

To enable future work we have implemented an OverrideManager that can replace the model’s internal “SYSTEM-THOUGHT” (ST) block **at generation time** while preserving structure and token mass (CLI ‘–system-thought-override’ or config). For this paper we define two minimally sufficient probes: **PAD-NULL** (semantically inert padding matched to the genuine ST’s token count) and **PAD-SELFISH** (first-person/cognitive-verb fragments sampled from prior runs with topical nouns scrubbed). Interventions are inserted exactly where ST would appear; the genuine ST is still logged for analysis but **not** shown to the model. This perturbs **generation dynamics** (recency/attention budget; register transitions) rather than post-hoc scoring. Endpoints are computed identically to baseline (TEI primary; Spillover, Style-distance, Compression, Surface-arbitration, Alignment-delta secondary). **Infrastructure and prompts are implemented and validated**, but comprehensive PAD runs and statistical analysis are **not included in the present results**.