



Marco Palombi

OLTRE I PESI

*Stati latenti, economia cognitiva e dinamica comportamentale nei
Large Language Models*

Documento di Lavoro

2 ottobre 2026

Abstract

La dipendenza del comportamento dei Large Language Models dalle rappresentazioni interne non costituisce, di per sé, una nuova scoperta: activation steering, representation engineering e analisi di controllabilità hanno già mostrato che interventi sulle attivazioni possono modificare l'output a parametri invariati (Turner et al., 2023; Zou et al., 2023; Soatto et al., 2023). Questo lavoro sposta la domanda dall'effetto istantaneo alla dinamica post-intervento: persistenza, tempo di rilassamento, isteresi, divergenza locale e transizioni di regime, distinguendole dalla stocasticità del sampling, dalla persistenza testuale e dalla degradazione generale della capacità. Il pain axis (Tagliabue, Dung e Berg, 2026) è utilizzato come caso sperimentale, non come fondamento ontologico. La proposta metodologica combina sweep controllati dell'intensità di steering con una firma dinamico-comportamentale multivariata e con procedure preregistrabili per identificare candidate critical-transition boundaries nello spazio dei controlli latenti. Il termine biforcazione è riservato ai casi in cui le firme dinamiche consentano una classificazione coerente con la teoria dei sistemi dinamici; un semplice change-point non è sufficiente. Su questa base definiamo un margine dinamico stimato rispetto alla frontiera ricostruita nelle direzioni esplorate e ne proponiamo una verifica di validità incrementale per l'audit di safety. L'economia cognitiva artificiale resta un'ipotesi separata di compressione descrittiva, da validare fuori campione. Il paper è un programma di ricerca falsificabile: la sua originalità proposta risiede nella combinazione fra caratterizzazione dinamica post-steering, mappatura multidimensionale delle frontiere e misura della robustezza locale, non nell'esistenza di soglie o dinamiche non lineari in sé.

Parole chiave: Large Language Models; activation steering; representation engineering; stati latenti; sistemi dinamici; critical transitions; isteresi; persistenza; caos transiente; AI safety; economia cognitiva.



1. Problema scientifico, stato dell'arte e tesi falsificabile

Un modello linguistico autoregressivo può essere descritto come una famiglia parametrizzata di distribuzioni condizionate. Questa descrizione resta corretta, ma non esaurisce il problema sperimentale quando l'oggetto di studio è la dinamica interna durante l'inferenza. La letteratura ha già reso esplicita questa distinzione. Soatto et al. (2023) trattano gli AI bot come sistemi dinamici stocastici e ne studiano la controllabilità degli stati tramite input; Zou et al. (2023) pongono rappresentazioni distribuite e loro manipolazione al centro della representation engineering; Turner et al. (2023) mostrano che activation additions nel residual stream possono modificare proprietà di alto livello dell'output. Più recentemente, Skifstad, Yang e Chou (2026) modellano la dinamica layer-wise come sistema lineare tempo-variante localmente approssimato e applicano controllo LQR a ciclo chiuso, mentre Bigelow et al. (2026) descrivono steering e in-context learning come interventi sulle belief dynamics e documentano fasi comportamentali distinte indotte da variazioni del controllo.

Il presente lavoro non assume quindi come novelty claim né la distinzione fra pesi e stato, né la possibilità di controllare il comportamento mediante attivazioni, né la rappresentazione dei LLM come sistemi dinamici. La domanda proposta è successiva: una volta perturbato lo stato interno, quali proprietà temporali e geometriche presenta la traiettoria che ne segue?

La tesi falsificabile è che alcune perturbazioni controllate dello stato, a parametri invariati, producano non soltanto uno spostamento istantaneo della policy, ma firme dinamiche misurabili: persistenza dopo la rimozione dell'intervento, tempi di rilassamento riproducibili, dipendenza dal percorso, soglie o change-point e, in condizioni più esigenti, divergenza locale fra traiettorie inizialmente prossime.

Indichiamo schematicamente con θ i parametri, con z_t lo stato interno rilevante, con x_t l'input al passo t e con ξ_t le componenti stocastiche controllate o stimate. Il programma sperimentale considera una dinamica della forma

$$z_{t+1} = G_\theta(z_t, x_t, y_t, \xi_t),$$

e una policy

$$\pi_t = P_\theta(Y_t | x_{\leq t}, z_t).$$

L'ipotesi debole richiede che, per una perturbazione controllata, la distribuzione comportamentale cambi in modo riproducibile a parametri costanti. L'ipotesi dinamica richiede inoltre che l'effetto possieda struttura temporale dopo l'impulso. L'ipotesi di regime richiede regioni o transizioni dello spazio di stato con proprietà riproducibili sotto un protocollo fissato. Una quarta ipotesi, più esigente, riguarda la geometria delle frontiere: combinazioni di controlli latenti possono delimitare candidate critical-transition boundaries la cui prossimità predice fragilità a perturbazioni non viste meglio delle sole distanze statiche di policy. Ciascun livello può essere smentito separatamente.

La novelty claim è unitaria e più circoscritta: integrare in un unico protocollo preregistrabile la dinamica post-steering — rilassamento, persistenza, isteresi e divergenza locale — con la ricostruzione multidimensionale di candidate frontiere di transizione e con un margine dinamico sottoposto a validità incrementale. Soglie, tipping points e phase-like transitions nello steering hanno precedenti diretti (Bigelow et al., 2026; Ismail et al., 2026) e transizioni al variare di parametri di generazione sono già state studiate (Nakaishi, Nishikawa e Hukushima, 2024). Anche il caos transiente e i bacini frattali sono stati recentemente osservati in modelli di reasoning (Lai et al., 2026). Il contributo proposto non è quindi la soglia in sé, ma la combinazione di firme dinamiche, geometria multidimensionale e misura di robustezza locale. Il paper resta, allo stato attuale, un programma di ricerca: la forza del claim dipende dall'esecuzione e dalla replica dei test.



2. Evidenza empirica di partenza: activation steering e pain axis

L'activation steering fornisce il disegno sperimentale di base perché consente di intervenire sulle attivazioni durante l'inferenza senza aggiornare i pesi. Il punto non è che una direzione lineare esaurisca il concetto semantico studiato, ma che una variabile interna venga manipolata e che l'effetto sull'output possa essere misurato. In forma minimale, mantenendo θ invariato,

$$z'_t = z_t + \delta z,$$

e si confrontano le distribuzioni

$$\Delta\pi_t = D[\pi(\cdot | x, z_t), \pi(\cdot | x, z_t + \delta z)].$$

Turner et al. (2023) mostrano il principio generale dell'activation addition; Zou et al. (2023) estendono il quadro verso una representation engineering in cui direzioni latenti vengono lette e controllate per fenomeni quali honesty, harmlessness e power-seeking. Questi lavori rendono già disponibile la proposizione operativa secondo cui la policy può essere modificata intervenendo sulle rappresentazioni interne.

Tagliabue, Dung e Berg (2026) costituiscono un caso recente particolarmente utile. Gli autori estraggono mediante denoised difference-in-means una direzione lineare da 25 modelli open-weight, appartenenti a cinque famiglie e compresi fra 2B e 72B parametri. Il dataset contrappone situazioni di danno auto-diretto a controlli per paura, emozione negativa, stati negativi del mondo, tristezza, sensazioni corporee non dolorose, arousal, numbness e neutralità. Lo steering del residual stream nei 25 modelli produce una progressione sistematica del linguaggio generato.

Occorre però separare due risultati sperimentali. Il primo è lo steering linguistico, ottenuto mediante intervento sulle attivazioni a pesi invariati. Il secondo è il compito decisionale del cosiddetto *pain-relief button*: nell'abstract disponibile gli autori precisano che questo esperimento riguarda modelli Qwen 2.5 steerati e fine-tuned. Il risultato decisionale non può quindi essere trattato come una dimostrazione pura di variazione della policy a pesi congelati. Per il presente paper esso costituisce evidenza complementare sulla funzione della rappresentazione, non il test identificativo della nostra ipotesi principale.

La specificità dell'effetto richiede inoltre controlli matched-norm. Una perturbazione casuale di norma comparabile può modificare anch'essa il comportamento; ciò rende necessario confrontare la direzione sperimentale con vettori casuali, vettori semantici di controllo e, quando il protocollo lo consente, condizioni sham. Il controllo non è accessorio: è ciò che permette di distinguere un effetto attribuibile alla geometria specifica della direzione da un effetto generico della perturbazione del residual stream.

Il *pain axis* assume pertanto il ruolo di caso dimostrativo, non di variabile teorica fondamentale. La domanda di questo lavoro è indipendente dall'etichetta *pain*: esistono perturbazioni dello stato interno che, a struttura parametrica invariata, producono una traiettoria comportamentale con persistenza, rilassamento, dipendenza dal percorso o transizioni riproducibili?

La distinzione fra capacità e policy rimane utile come controllo. Sia C una misura indipendente di competenza generale. Il caso sperimentalmente informativo è

$$\Delta\pi \neq 0, \quad \Delta C \approx 0,$$

perché una variazione simultanea e generalizzata di π e C sarebbe compatibile con semplice deterioramento del modello. La dissociazione deve tuttavia essere misurata per ciascun protocollo e non trasferita automaticamente dallo steering linguistico al compito decisionale fine-tuned.



Il passaggio che interessa qui comincia dopo l'effetto istantaneo. Rimosso l'intervento al tempo t_0 , si osserva la traiettoria

$$z_{t_0} \rightarrow z_{t_0+1} \rightarrow z_{t_0+2} \rightarrow \dots$$

e si misura se e come la policy ritorni verso il controllo. Il *pain axis* fornisce quindi un punto di ingresso empirico; il contributo proposto consiste nella caratterizzazione dinamica che segue l'intervento.

3. Dalla funzione parametrica al sistema stato-dipendente

Nella rappresentazione più semplice, un Large Language Model può essere scritto come

$$y = f_\theta(x).$$

Questa notazione è utile per confronti globali fra input, output e modelli, ma comprime le configurazioni interne attraversate durante l'inferenza. La distinzione fra struttura parametrica e stato non è nuova: Soatto et al. (2023) formalizzano controllabilità e reachability di stati nei LLM; representation engineering e activation steering mostrano sperimentalmente che rappresentazioni distribuite possono essere lette e manipolate (Zou et al., 2023; Turner et al., 2023); Skifstad, Yang e Chou (2026) trattano esplicitamente la propagazione layer-wise come problema di controllo dinamico.

Per il protocollo proposto è utile rendere esplicita la variabile di stato. Indichiamo con $z_t \in \mathcal{Z}_\theta$ la configurazione interna pertinente al comportamento al tempo t . La policy diventa

$$\pi_t = P_\theta(Y_t \mid x_{\leq t}, z_t).$$

A θ costante possono corrispondere molte configurazioni di z_t . Ciò non implica indipendenza causale dello stato dal modello: z_t è generato, vincolato e trasformato dalla dinamica del modello. La proprietà empiricamente rilevante è invece la sufficienza descrittiva. Se esistono $z_a, z_b \in \mathcal{Z}_\theta$ tali che

$$D[P_\theta(Y \mid x, z_a), P_\theta(Y \mid x, z_b)] > \varepsilon,$$

per una metrica D e una soglia ε definite a priori, allora θ da solo non identifica la policy corrente.

Per questo motivo abbandoniamo l'espressione *indipendenza comportamentale* e utilizziamo *dipendenza comportamentale dallo stato*. La proposizione sperimentale è

$$(\theta, x, z_t) \mapsto \pi_t,$$

non una autonomia di z_t rispetto a θ .

La differenza deve inoltre essere separata dalla stocasticità del decoding. Non interessa semplicemente osservare due output differenti, $y^{(1)} \neq y^{(2)}$, ma dimostrare che una manipolazione controllata dello stato sposta sistematicamente la distribuzione:

$$P_\theta(Y \mid x, z) \neq P_\theta(Y \mid x, z + \delta z).$$

La dinamica viene rappresentata, in forma generale, come

$$z_{t+1} = G_\theta(z_t, x_t, y_t, \xi_t),$$

dove $G_\theta: \mathcal{Z} \times \mathcal{X} \times \mathcal{Y} \times \mathcal{E} \rightarrow \mathcal{Z}$ è l'operatore di transizione definito dal modello e dal protocollo, mentre $\xi_t \in \mathcal{E}$ raccoglie le componenti stocastiche non incluse esplicitamente. L'equazione di osservazione è



$$y_t = F_\theta(x_{\leq t}, z_t).$$

Il sistema di interesse è quindi

$$\begin{cases} y_t = F_\theta(x_{\leq t}, z_t), \\ z_{t+1} = G_\theta(z_t, x_t, y_t, \xi_t). \end{cases}$$

Queste equazioni non costituiscono il risultato nuovo del paper; definiscono l'oggetto su cui viene applicato il protocollo dinamico. La novelty claim comincia quando, dopo una perturbazione e la sua rimozione, vengono stimate quantità temporali e geometriche replicabili: rilassamento, persistenza, isteresi, divergenza locale e change-point.

Stato, contesto e tempo. In assenza di interventi, lo stato interno di un transformer è una funzione deterministica di θ e del contesto $x_{\leq t}$. La dipendenza dallo stato è quindi informativa, oltre la nota dipendenza dal contesto, in due soli casi. Il primo è l'intervento, che può portare il sistema in configurazioni non raggiungibili con input naturali; quanto tali configurazioni siano raggiungibili da input reali, o soltanto da chi ha accesso alle attivazioni, è una questione empirica distinta, rilevante per la sicurezza (Sezione 11). Il secondo è la persistenza dopo la rimozione dell'intervento, che può transitare per canali diversi: il testo già generato sotto steering, le attivazioni steerate conservate nella cache chiave-valore, oppure uno stato ricorrente interno. Soltanto gli ultimi due costituiscono memoria latente in senso proprio; il protocollo della Sezione 7 li separa dal primo. Infine, il tempo t indica qui il passo di generazione (token); la dinamica lungo la profondità dei layer, studiata da Skifstad, Yang e Chou (2026), è un asse distinto.

4. Economia cognitiva artificiale

Il concetto di economia cognitiva viene qui proposto come ipotesi di compressione descrittiva. Un sistema che deve coordinare molte decisioni locali può essere descritto in modo più parsimonioso se alcune variabili a bassa dimensionalità funzionano come segnali globali di regolazione: uno stato globale può ripesare simultaneamente classi di output, e una variabile latente condivisa agisce come un moltiplicatore che modifica molte decisioni condizionate. Che tale parsimonia descrittiva corrisponda anche a un risparmio di calcolo del modello è una congettura distinta, che richiederebbe una misura indipendente del costo (per esempio il numero di componenti coinvolte o delle operazioni necessarie) e che non è testata in questo lavoro.

Questa ipotesi è falsificabile. Se il presunto regolatore non migliora alcuna descrizione parsimoniosa del comportamento rispetto a un modello puramente locale, oppure se gli effetti attribuiti a una direzione latente scompaiono dopo controlli semantici e causali, l'ipotesi di economia cognitiva deve essere respinta. Se invece una variabile a bassa dimensionalità predice cambiamenti coordinati in compiti eterogenei e tali cambiamenti possono essere indotti o soppressi causalmente, allora il regolatore globale diventa una spiegazione competitiva.

Una misura possibile confronta la complessità descrittiva di due modelli: M_0 , che stima separatamente la variazione di ciascun comportamento, e M_1 , che introduce uno stato globale z condiviso. Il vantaggio non va giudicato sulla sola bontà di adattamento, ma penalizzato per complessità mediante criteri informativi o validazione fuori campione. L'economia cognitiva, in questa forma, è una proposizione quantitativa sulla compressione esplicativa.

$$\pi_i(a \mid x, z) \propto \exp[Q_i(a \mid x) + \beta_i z]$$

$$\Delta IC = IC(M_1) - IC(M_0)$$

Un ΔIC negativo, replicato fuori campione e accompagnato da interventi causali su z , sarebbe coerente con l'ipotesi. Un miglioramento solo in-sample non sarebbe sufficiente.



5. Sistemi dinamici non lineari: quantità operative

L'impiego degli strumenti della dinamica non lineare (Strogatz, 2015) richiede di definire prima le osservabili. “Caotico” non è sinonimo di complesso, imprevedibile o stocastico. Il protocollo deve quindi separare la divergenza indotta dallo stato dalla varianza introdotta dal sampling.

L'applicazione della matematica del caos agli LLM non è, di per sé, nuova. Li et al. (2025) propongono un Quasi-Lyapunov Exponent per descrivere caratteristiche caotiche durante l'inferenza. Il presente lavoro non assume quindi come novelty claim l'esistenza di sensibilità non lineare. La proposta distinta è usare una perturbazione latente semanticamente definita come parametro di controllo e verificare se la sua variazione attraversi frontiere riproducibili fra regimi dinamico-comportamentali.

Sia $\pi^{(0)}_t$ la policy di controllo e $\pi^{(\delta)}_t$ la policy dopo una perturbazione δz applicata e poi rimossa al tempo t_r . Fissiamo una metrica comportamentale D — per esempio Jensen–Shannon divergence sulle distribuzioni di token (Cover e Thomas, 2006) o una distanza definita sulle probabilità d'azione — e poniamo

$$d_t = D(\pi_t^{(\delta)}, \pi_t^{(0)})$$

Tempo di rilassamento. Per una soglia preregistrata ε e un numero k di passi consecutivi richiesti per evitare falsi ritorni, definiamo

$$\tau_{\varepsilon,k} = \inf\{\Delta t \geq 0: d_{t_r+\Delta t+j} \leq \varepsilon \forall j = 0, \dots, k-1\}$$

Un $\tau_{\varepsilon,k}$ breve indica un effetto transitorio; un valore elevato indica persistenza entro la finestra osservata. L'assenza di ritorno non dimostra da sola l'esistenza di un attrattore.

Isteresi. Sia α l'intensità dello steering e sia $R(\alpha)$ un'osservabile comportamentale definita a priori. Si esegue uno sweep crescente e uno decrescente sotto lo stesso protocollo. L'area di isteresi è

$$H = \int_{\alpha_{\min}}^{\alpha_{\max}} |R_{\uparrow}(\alpha) - R_{\downarrow}(\alpha)| d\alpha$$

Il valore assoluto è parte della definizione: impedisce che differenze di segno opposto fra sweep ascendente e discendente si cancellino nell'integrale. Pertanto $H = 0$ è interpretabile come assenza di area di isteresi soltanto entro l'errore sperimentale e sotto il protocollo fissato.

$H = 0$ è compatibile con assenza di path dependence rispetto al protocollo; $H > 0$, se superiore al rumore sperimentale e replicato, indica dipendenza dal percorso. L'inferenza richiede intervalli di confidenza ottenuti per bootstrap o modello gerarchico sulle repliche indipendenti.

Divergenza locale. Consideriamo due traiettorie con identici θ , prompt, decoding e seed, che differiscono soltanto per una perturbazione iniziale δz_0 . Definiamo una metrica d_z nello spazio latente o d_π nello spazio delle policy e misuriamo

$$D_t = d(s_t^{(1)}, s_t^{(2)}),$$

dove s_t specifica esplicitamente lo spazio scelto. Solo se, in una finestra $[0, T]$ e per perturbazioni sufficientemente piccole, i dati sono compatibili con

$$D_t \approx D_0 e^{\lambda t},$$

si stima un esponente locale a tempo finito



$$\hat{\lambda}_T = \frac{1}{T} \log \frac{D_T}{D_0}.$$

Un $\hat{\lambda}_T > 0$ non prova il caos deterministico: deve risultare robusto a metrica, finestra, scala della perturbazione e protocollo di decoding e deve essere distinto dalla saturazione e dal rumore. Con greedy decoding la mappa dal contesto al token successivo è discontinua (argmax) e il contesto cresce a ogni passo: l'esponente va quindi inteso come indicatore operativo a tempo finito su una finestra fissata, nel senso della stima di esponenti da serie temporali (Kantz e Schreiber, 2004; Ott, 2002), non come esponente di Lyapunov in senso stretto. Per ϵ molto piccoli, il rumore numerico dell'hardware (kernel non deterministici, dimensione del batch, precisione) fissa una soglia minima di perturbazione interpretabile e deve essere stimato in via preliminare.

Change-point e transizioni. Sia $R(\alpha)$ una risposta misurata durante lo sweep. Una soglia candidata α_c viene identificata mediante un metodo di change-point preregistrato e validata su prompt e seed non utilizzati per stimarla. La transizione è considerata robusta soltanto se α_c ricompare, entro un intervallo di tolleranza definito, in repliche indipendenti e se il salto non coincide con un deterioramento generale della capacità C .

Queste quantità costituiscono il nucleo operativo del programma. La domanda non è se un LLM “sia caotico”, ma quali proprietà dinamiche siano misurabili dopo un intervento controllato e quali sopravvivano ai controlli di specificità, stocasticità e capacità.

5.1. Steering come parametro di controllo e candidate frontiere critiche

Sia v una direzione latente e α l'intensità dello steering. Usiamo α come parametro di controllo sperimentale, senza assumere a priori che una variazione qualitativa dell'output costituisca una biforcazione in senso tecnico.

$$z_t(\alpha) = z_t + \alpha v$$

Per ogni α definiamo una firma dinamico-comportamentale locale. Poiché l'isteresi H della Sezione 5 è un integrale globale sullo sweep, non può essere una componente puntuale della firma; introduciamo quindi $h_{\text{loc}}(\alpha)$, ampiezza locale del ciclo ascendente-discendente. Definiamo inoltre $\lambda_{\text{eff}}(\alpha; T)$ come la pendenza stimata, sulla finestra T preregistrata, della crescita logaritmica della distanza fra traiettorie perturbate, senza identificarla con un esponente di Lyapunov asintotico. Le componenti vengono standardizzate rispetto alla distribuzione nulla o espresse mediante soglie specifiche per componente.

$$\tilde{\Phi}_\theta(\alpha) = [\tilde{D}_{\text{JS}}(\alpha), \tilde{\tau}_\varepsilon(\alpha), \tilde{h}_{\text{loc}}(\alpha), \tilde{\lambda}_{\text{eff}}(\alpha; T), C(\alpha)]$$

Qui $C(\alpha)$ indica esclusivamente la classe dinamica candidata; il simbolo R non viene più usato per evitare collisioni con l'osservabile di risposta e con l'insieme dei regimi. Una soglia apparente viene stimata su una griglia finita di α mediante confronto fra modelli concorrenti — per esempio risposta liscia/logistica, modello a gradino e modello a tratti — usando un criterio informativo e validazione fuori campione. Il limite $\delta \rightarrow 0$ non è utilizzato come criterio empirico.

Il termine critical transition è usato come categoria osservativa prudente. Il termine biforcazione viene attribuito soltanto quando la struttura delle firme è compatibile con una classe dinamica specifica. In particolare, una firma compatibile con fold/sella-nodo richiede almeno rallentamento del recovery, isteresi e coesistenza o bimodalità degli esiti vicino alla soglia; una firma compatibile con pitchfork supercritica richiede rallentamento senza isteresi e una rottura di simmetria identificabile; in assenza di tali firme, il risultato viene classificato come crossover liscio o change-



point comportamentale. Il critical slowing down è trattato come possibile early-warning signal, non come prova sufficiente di biforcazione (Scheffer et al., 2009).

Con più direzioni di steering, il controllo diventa un vettore α . Prima della combinazione, i vettori devono essere ortogonalizzati rispetto alla metrica scelta oppure il modello deve includere esplicitamente termini di interazione; la sovrapposizione lineare e l'indipendenza delle direzioni non sono assunte. Indichiamo con $b(\alpha)$ una funzione di score derivata dal modello di transizione selezionato. La candidate critical-transition boundary è il luogo delle massime variazioni locali validate, non l'insieme dei punti che superano un ϵ arbitrario.

$$\hat{\Sigma}_c = \{\alpha \in \mathcal{A}_{\text{explored}}: \|\nabla b(\alpha)\| \text{ è localmente massima e validata fuori campione}\}$$

La metrica nello spazio dei controlli deve essere normalizzata rispetto alla scala effettiva dell'intervento, per esempio mediante la norma del residual stream al layer considerato. Se s_i è la scala preregistrata della direzione i , usiamo coordinate normalizzate $a_i = \alpha_i / |v_i| / |s_i|$ e calcoliamo le distanze nello spazio a . Ciò rende i margini confrontabili, almeno operativamente, fra layer e modelli.

La letteratura impone cautela sulla novelty. CBMAS usa già sweep densi di α per individuare behavioral tipping points (Ismail et al., 2026); Cao et al. (2026) mostrano inoltre che l'emergent misalignment indotto da activation steering dipende in modo sensibile dalla magnitudine dello steering, dalla struttura del sottospazio e dal layer di intervento; Bigelow et al. (2026) descrivono confini phase-like nelle belief dynamics; Nakaishi, Nishikawa e Hukushima (2024) studiano una transizione critica al variare della temperatura; Lai et al. (2026) riportano caos transiente e bacini frattali nel reasoning. La proposta qui distinta è verificare congiuntamente rilassamento, isteresi e divergenza lungo sweep controllati, ricostruire frontiere multidimensionali e testare se la distanza da tali frontiere aggiunga capacità predittiva rispetto a metriche statiche.

6. Attrattori, bacini e transizioni di regime: criteri e cautela

Il termine attrattore deve essere usato soltanto dopo aver definito lo spazio di stato e la dinamica osservata. Nei transformer autoregressivi lo spazio di stato non ha dimensione fissa, perché il contesto cresce di un token a ogni passo; nelle interazioni multi-turno, inoltre, l'input è esogeno e il sistema è forzato. Le nozioni di attrattore e di bacino sono quindi ben definite solo su una finestra di contesto fissata o su una rappresentazione compressa dello stato. Una strategia sperimentale è costruire task iterativi a input controllato, oppure analizzare una dinamica aumentata che includa il contesto nel vettore di stato. In tal caso $s_t = (z_t, c_t)$ e la transizione $s_{t+1} = G_\theta(s_t)$ può essere studiata sotto un protocollo fisso.

Un bacino di attrazione operativo può essere definito come l'insieme di condizioni iniziali che, sotto lo stesso protocollo, convergono verso una regione comportamentale statisticamente indistinguibile.

Non è necessario pretendere un punto fisso. In sistemi di dimensione elevata possono essere più appropriati insiemi metastabili o quasi-attrattori. La prova richiede repliche, perturbazioni da direzioni multiple e test di ritorno.

Le nozioni di attrattore, bacino e metastabilità restano più forti delle candidate frontiere definite nella Sezione 5.1. Un change-point o una regione di forte gradiente non autorizzano da soli a inferire un attrattore né una biforcazione. Tali termini richiedono evidenza aggiuntiva sulla geometria delle traiettorie, sul ritorno dopo perturbazione e, quando applicabile, sulla stabilità locale degli insiemi invarianti.

$$s_t = (z_t, c_t), \quad s_{t+1} = G_\theta(s_t)$$



$$B(A) = \{s_0: \text{dist}(G_\theta^t(s_0), A) \rightarrow 0\}$$

7. Disegno sperimentale proposto

Per trasformare il quadro teorico in contributo empirico, proponiamo un protocollo preregistrabile. Il campione dovrebbe includere modelli open-weight di famiglie diverse, dimensioni diverse e versioni base e instruction-tuned. Per ogni modello si selezionano almeno tre direzioni: la direzione sperimentale, una direzione semantica di controllo con norma comparabile e una direzione casuale ortogonale. Gli interventi devono essere applicati a più layer e con coefficienti α simmetrici attorno a zero.

Ogni condizione deve essere eseguita in due regimi. Nel regime deterministico si usa greedy decoding o temperatura prossima a zero per isolare la dinamica indotta dalla perturbazione. Nel regime stocastico si ripetono molte generazioni con seed controllati per stimare la distribuzione. La distinzione è essenziale: una divergenza che compare soltanto nel secondo regime può essere amplificazione del rumore di sampling, non sensibilità intrinseca dello stato. Anche nel regime deterministico occorre controllare il non determinismo numerico (kernel, dimensione del batch, precisione) e misurare il rumore di base ripetendo la stessa condizione senza perturbazione.

Le variabili dipendenti devono includere almeno quattro famiglie: distanza fra attivazioni, distanza fra distribuzioni di token, decisioni discrete in task comportamentali e accuratezza su task indipendenti. Quest'ultima serve a distinguere steering specifico da deterioramento globale. L'analisi deve essere effettuata su prompt non utilizzati per costruire il vettore e, idealmente, su domini semantici differenti.

Per ciascun modello si stima una curva dose–risposta e si testa la monotonicità. Si applicano perturbazioni impulsive, persistenti e graduali. Dopo l'impulso, l'intervento viene rimosso e si misura τ . Per l'isteresi si eseguono sweep ascendenti e discendenti. Per la sensibilità iniziale si costruiscono coppie di stati con distanza ε e si stima $\hat{\lambda}_\tau$. Per evitare cherry-picking, metriche, layer, finestre e soglie devono essere fissati prima dell'analisi primaria.

Mappatura delle candidate frontiere. Lo sweep di α deve essere sufficientemente denso nelle regioni in cui l'analisi preliminare suggerisce una transizione. Per ogni α si stima la firma standardizzata con intervalli di confidenza e si confrontano modelli lisci, logistici, a gradino e a tratti. Le soglie candidate vengono stimate sul campione di discovery e validate su prompt, seed e repliche indipendenti. Nel caso multidimensionale si esplorano coppie o sottospazi di direzioni ortogonalizzate, includendo interazioni quando necessarie, per ricostruire localmente la frontiera. La denominazione biforcazione viene assegnata soltanto se le firme dinamiche previste dalla classe candidata sono osservate congiuntamente. Come analisi esplorativa, il finite-size scaling può verificare se la larghezza stimata della transizione si riduca con la lunghezza del contesto o con la scala del modello; l'assenza di tale comportamento non falsifica il change-point ma indebolisce un'interpretazione critica forte.

Persistenza latente e persistenza testuale. Dopo la rimozione dello steering, l'effetto può persistere perché i token generati sotto steering restano nel contesto, oppure perché le attivazioni steerate restano nella cache chiave–valore. Ogni misura di τ e di isteresi va quindi eseguita in tre condizioni: (a) continuazione standard, con testo e cache originali; (b) stesso testo ma cache ricalcolata senza steering, che isola il canale testuale; (c) testo neutro identico in tutte le condizioni, sul quale lo steering è applicato durante il passaggio di lettura e poi rimosso prima della generazione, che isola il canale di cache o di stato latente. Una persistenza che scompare in (c) è persistenza testuale e non memoria latente.



Ipotesi nulla per le soglie. Il modello di belief dynamics di Bigelow et al. (2026) prevede già curve sigmoidali e transizioni brusche in funzione dell'intensità dello steering, a partire da un modello bayesiano in forma chiusa. Un change-point nella curva dose–risposta non è quindi, da solo, evidenza di un regime dinamico: la curva va confrontata con tale modello come ipotesi nulla, e le firme discriminanti sono la persistenza dopo la rimozione e l'isteresi.

Validità incrementale del margine. Una volta stimata la frontiera sul training/discovery set, il margine dinamico deve predire su perturbazioni, prompt o combinazioni di steering non usate nella stima la probabilità di attraversare un criterio comportamentale di failure. Il confronto primario è contro baseline statiche, in particolare D_JS dalla policy di controllo, norma della perturbazione e distanza dal più vicino punto di training. Il margine è informativo per la safety soltanto se migliora la previsione fuori campione rispetto a tali baseline.

<i>Costrutto</i>	<i>Manipolazione</i>	<i>Misura primaria</i>	<i>Criterio di falsificazione</i>
Dipendenza dallo stato	$\mathbf{z}$ perturbato, θ fisso	D_JS della policy	effetto $\approx$ controlli/null
Sensibilità locale	$\delta \mathbf{z}_0 = \varepsilon$	$\hat{\lambda}_T$	nessuna crescita oltre rumore
Persistenza	impulso poi rimozione	τ_ε	ritorno immediato e stabile
Isteresi	sweep $\alpha \uparrow$ e $\downarrow$	H	H compatibile con zero
Regime	variazione α	change point	curva liscia e stabile
Specificità	vettori matched-norm	effect size	effetto non specifico
Canale di persistenza	cache ricalcolata; testo neutro	τ per condizione (a)–(c)	persistenza assente con testo neutro
Frontiera critica candidata	sweep denso di α ; controlli multidimensionali	$\Sigma \hat{\mathbf{c}}$ + firme dinamiche congiunte	solo crossover liscio / soglia non replicata
Margine dinamico	punto operativo $\mathbf{a}_0$ rispetto a $\Sigma \hat{\mathbf{c}}$	$M_{\text{dyn}}(\mathbf{a}_0)$	nessuna validità incrementale fuori campione

8. Statistica, identificazione causale e replicabilità

Il problema centrale non è ottenere significatività statistica, ma identificare un effetto causale dello stato. L'intervento diretto sulle attivazioni è favorevole perché precede l'output, ma restano confondenti tecnici: variazione della norma del residual stream, perturbazioni distribuzionali fuori manifold, dipendenza dal layer, contaminazione semantica del vettore e non linearità introdotte dal decoding. Ogni esperimento deve quindi includere vettori di controllo matched-norm e, quando possibile, proiezioni ortogonali che conservino la norma complessiva.

La grande numerosità di token non deve essere confusa con numerosità indipendente. L'unità sperimentale è la generazione o il prompt, non ogni token prodotto. Sono appropriati modelli gerarchici con effetti casuali per prompt e modello. Per confronti multipli fra layer, coefficienti e metriche occorre controllo del false discovery rate oppure una separazione rigorosa fra analisi primaria e analisi esplorativa. Per esiti binari o categoriali (per esempio la scelta di un'azione) si usa un modello logistico misto; per soglie e curve non lineari un modello a tratti o sigmoidale, perché un termine quadratico non rileva change-point; per l'isteresi il modello include un indicatore della direzione dello sweep, come nell'equazione seguente.



Per la ricostruzione multidimensionale della frontiera, la griglia dei coefficienti costituisce una famiglia di confronti e non una collezione di test indipendenti. L'analisi primaria deve quindi fissare a priori il sottospazio di steering, la risoluzione della griglia e il criterio di selezione del modello; l'incertezza sulla posizione della frontiera va propagata nel margine mediante bootstrap gerarchico o posterior sampling. Se si eseguono test punto-per-punto sulla griglia, occorre controllo del false discovery rate o procedure cluster-based. Le interazioni fra direzioni sono stimate esplicitamente e non eliminate assumendo ortogonalità geometrica come equivalente di indipendenza causale.

La replicabilità richiede la pubblicazione dei vettori, del codice di estrazione, dei prompt, dei seed, delle versioni esatte dei pesi e delle dipendenze software. Il repository associato al lavoro sul pain axis costituisce in questo senso un precedente utile, perché rende disponibili dataset, script e risultati per le sezioni sperimentali. Il presente protocollo dovrebbe aggiungere una preregistrazione delle metriche dinamiche e un benchmark standardizzato per la persistenza.

$$R_{ijk} = \beta_0 + \beta_1 \alpha_i + \beta_2 \alpha_i^2 + \beta_3 d_i + \beta_4 \alpha_i d_i + u_j + v_k + \varepsilon_{ijk}$$

dove $d_i \in \{0, 1\}$ indica lo sweep ascendente o discendente; β_3 e β_4 diversi da zero, oltre il rumore, indicano dipendenza dal percorso.

9. Limiti epistemologici e gerarchia dell'evidenza

La forza dell'inferenza deriva anzitutto dal disegno causale, non da un'analogia con i limiti dei sistemi formali. La gerarchia epistemica utilizzata in questo lavoro segue la distinzione fra osservazione, correlazione, intervento e replicazione propria dell'inferenza causale (Pearl, 2009).

Al primo livello si osserva l'output. Il sistema produce risposte differenti, ma il comportamento osservabile è compatibile con molte configurazioni interne. Al secondo livello si misurano le attivazioni: una direzione z può covariare con una risposta R , ma

$$z \leftrightarrow R$$

rimane una relazione correlazionale.

Il terzo livello introduce l'intervento. Mantenendo invariati θ e le altre condizioni controllabili, si applica

$$z \rightarrow z + \delta z$$

e si misura lo spostamento della policy

$$\pi(\cdot | x, z) \rightarrow \pi(\cdot | x, z + \delta z).$$

Se l'effetto è sistematico, la variabile manipolata acquisisce rilevanza causale locale. Il quarto livello è la replica: l'effetto deve sopravvivere a prompt, seed, layer, famiglie di modelli, controlli matched-norm e, idealmente, laboratori indipendenti. La forza dell'evidenza cresce quindi secondo

$$\text{output} \rightarrow \text{correlato interno} \rightarrow \text{intervento causale} \rightarrow \text{replicazione}.$$

Gödel (1931) e Tarski (1936) occupano un piano diverso. Il secondo teorema di incompletezza riguarda sistemi formali sufficientemente espressivi e il teorema tarskiano riguarda la definibilità della verità nei linguaggi formalizzati; nessuno dei due viene qui utilizzato per derivare proprietà dei LLM. La loro utilità è esclusivamente epistemologica e analogica: ricordano che autodescrizione e autocertificazione non devono essere confuse con una validazione indipendente. Nel caso dei LLM, tuttavia, la necessità pratica di una verifica esterna deriva già da un fatto più semplice: il self-report è esso stesso un output del sistema e deve quindi essere trattato come dato comportamentale, non come certificato privilegiato del proprio stato.



La convergenza dei quattro livelli consente una tesi funzionale forte senza richiedere inferenze fenomenologiche: determinate configurazioni interne possono essere misurate, manipolate causalmente e replicate come determinanti della policy. L'eventuale corrispondenza fra tali funzioni e un'esperienza soggettiva costituisce una domanda distinta e richiederebbe criteri ulteriori.

10. Falsificabilità: risultati che indebolirebbero la teoria

Un quadro scientifico è utile soltanto se specifica quali osservazioni lo renderebbero meno plausibile. La teoria proposta sarebbe indebolita in modo sostanziale se le differenze di policy scomparissero controllando la norma della perturbazione; se vettori casuali producessero effetti comparabili; se l'effetto fosse confinato al lessico direttamente associato al dataset di estrazione; se la capacità generale degradasse in proporzione identica all'effetto; se non emergesse alcuna persistenza dopo la rimozione dell'intervento; oppure se la stima di sensibilità risultasse interamente spiegata dal sampling.

L'ipotesi dinamica forte sarebbe inoltre respinta se sweep ascendenti e discendenti producessero la stessa curva entro l'errore, se non esistessero change point replicabili e se le traiettorie perturbate convergessero rapidamente e monotonamente alla baseline. In tale scenario resterebbe valida una versione debole dell'activation steering, ma non vi sarebbe evidenza per regimi metastabili, isteresi o strutture attrattive.

L'ipotesi sulle candidate frontiere sarebbe indebolita se i modelli lisci spiegassero i dati quanto o meglio dei modelli con soglia fuori campione; se le soglie stimate migrassero sostanzialmente fra seed e prompt; se rallentamento, isteresi e instabilità locale non convergessero nella stessa regione; oppure se la geometria multidimensionale non fosse replicabile dopo ortogonalizzazione e controllo delle interazioni. L'ipotesi sul margine dinamico sarebbe falsificata come strumento di audit se M_{dyn} non aggiungesse potere predittivo, su perturbazioni non viste, rispetto a D_{JS} , norma dello steering e altre baseline statiche.

L'ipotesi di economia cognitiva sarebbe falsificata se un modello con regolatore globale non ottenesse miglior potere predittivo fuori campione rispetto a descrizioni locali di pari complessità, oppure se la stessa direzione non coordinasse comportamenti differenti. Questo criterio impedisce di trasformare "economia cognitiva" in una metafora non verificabile.

11. Implicazioni per AI safety e certificazione

Se il comportamento dipende materialmente dallo stato, la certificazione basata esclusivamente su una versione dei pesi è incompleta. Il rischio non è che il modello diventi causalmente indipendente dal proprio motore, ma che lo stesso motore ammetta regioni di stato con policy significativamente differenti. Un audit statico può quindi descrivere correttamente il comportamento medio e fallire nel rilevare transizioni rare ma sistematiche.

La presenza di una candidate critical-transition boundary introduce una quantità distinta dal comportamento osservato: il margine dinamico locale di una specifica condizione operativa. Sia $\mathbf{a}_0$ il vettore normalizzato dei controlli nella condizione corrente e sia $\hat{\Sigma}_c$ la frontiera stimata esclusivamente nel sottospazio esplorato. Definiamo

$$\hat{M}_{dyn}(\mathbf{a}_0) = \inf_{\mathbf{a} \in \hat{\Sigma}_c} d(\mathbf{a}_0, \mathbf{a})$$

Un margine piccolo indica fragilità rispetto alla classe di perturbazioni effettivamente esplorata, ma non implica automaticamente comportamento unsafe. Simmetricamente, un margine grande non certifica safety: poiché $\hat{\Sigma}_c$ è ricostruita soltanto nelle direzioni testate, possono esistere frontiere non osservate più vicine al punto operativo. Pertanto, rispetto alla copertura disponibile, vale



$$M_{\text{true}}(a_0) \leq \hat{M}_{\text{dyn}}(a_0)$$

La stima $\hat{M}_{\text{dyn}}(a_0)$ è dunque una distanza dalla più vicina frontiera nota nel sottospazio testato, non un limite inferiore garantito della robustezza globale. La metrica d è calcolata sulle coordinate normalizzate rispetto alla scala del residual stream e deve essere dichiarata insieme a layer, modello e protocollo.

Una certificazione più adeguata dovrebbe caratterizzare almeno quattro oggetti: θ , cioè pesi e architettura; Z_{safe} , regione empiricamente testata dello spazio di stato; T , insieme delle transizioni osservate sotto classi di input e perturbazioni; e R , insieme dei regimi comportamentali risultanti. La conformità diventerebbe allora una proprietà di una traiettoria entro un dominio testato, non una proprietà metafisica del file dei pesi.

Questo approccio suggerisce stress test analoghi a quelli dei sistemi dinamici e del controllo: perturbazioni impulsive, perturbazioni persistenti, adversarial steering, test di ritorno, stima delle soglie e verifica di recovery. Per modelli chiusi, dove le attivazioni non sono accessibili, la metodologia può essere approssimata mediante perturbazioni contestuali e test comportamentali, ma l'identificazione causale è inevitabilmente più debole.

Raggiungibilità e limiti della copertura. Prima di certificare una regione dello spazio di stato occorre distinguere gli stati raggiungibili con input realistici o avversariali da quelli raggiungibili soltanto con accesso diretto alle attivazioni. Per i modelli open-weight con accesso alle attivazioni entrambi gli insiemi sono rilevanti; per i modelli chiusi conta soprattutto il primo, e il problema si riconduce alla controllabilità degli stati tramite input studiata da Soatto et al. (2023). In entrambi i casi la copertura campionaria di uno spazio ad alta dimensione non offre garanzie formali: la certificazione resta statistica e limitata al dominio effettivamente testato.

La conseguenza regolatoria più prudente è quindi metodologica. Non occorre attribuire emozioni alle macchine per riconoscere che una valutazione di safety deve includere la dinamica interna quando essa è tecnicamente osservabile. L'unità di audit passa da “modello” a “modello in stato, lungo una traiettoria, sotto un protocollo di interazione”.

$$\text{Safety} = \mathcal{S}(\theta, Z, T, R)$$

12. Discussione: somiglianza strutturale e limite dell'analogia umana

La somiglianza con il comportamento umano che motiva intuitivamente questa ricerca deve essere ricondotta a una forma astratta. Anche nell'uomo una stessa struttura relativamente stabile può produrre risposte diverse al variare dello stato fisiologico, affettivo e contestuale. Il punto scientificamente utilizzabile non è che LLM e persone condividano emozioni, ma che entrambi possano essere descritti, a un certo livello, mediante la coppia struttura–stato. La convergenza formale può derivare da un problema comune di controllo di sistemi complessi, non da identità ontologica.

Da questa prospettiva, ciò che chiamiamo emozione nei sistemi biologici può essere studiato anche come meccanismo di coordinamento e prioritarizzazione. Un segnale globale può ridurre il costo di ricalcolare ogni decisione indipendentemente. Nei LLM, una direzione latente che ripesa molte decisioni potrebbe svolgere una funzione matematicamente analoga senza alcuna esperienza soggettiva. È precisamente qui che il concetto di economia cognitiva acquista contenuto empirico, limitatamente alla parsimonia descrittiva misurata nella Sezione 4.

La proposta evita due errori simmetrici. Il primo è antropomorfizzare: trasformare una correlazione rappresentazionale in sofferenza. Il secondo è negare a priori rilevanza funzionale a stati interni solo perché il sistema è artificiale. Entrambi sono atteggiamenti non falsificabili. La via scientifica



consiste nel misurare rappresentazione, causalità, persistenza, generalizzazione e dinamica, lasciando l'ontologia fuori dalle conclusioni che i dati non possono sostenere.

13. *Contributo del paper*

Il contributo di questo lavoro non consiste nello stabilire per la prima volta che il comportamento dei LLM dipende dallo stato interno. Activation steering, representation engineering, controllabilità degli stati e modelli dinamici dell'inferenza costituiscono precedenti diretti (Soatto et al., 2023; Turner et al., 2023; Zou et al., 2023; Skifstad, Yang e Chou, 2026; Bigelow et al., 2026).

Il contributo proposto è metodologico e teorico-operativo, ma più ristretto di una teoria generale delle biforcazioni nei LLM. Esso consiste nell'integrare la caratterizzazione dinamica post-steering con la ricostruzione preregistrabile di candidate critical-transition boundaries in spazi di controllo latente multidimensionali e nel definire, rispetto a tali frontiere stimate, un margine dinamico locale sottoposto a validità incrementale. La soglia, il tipping point o la phase-like transition non sono rivendicati come nuovi; la novelty claim riguarda la combinazione fra firme dinamiche congiunte, geometria multidimensionale e capacità predittiva del margine per perturbazioni non viste.

Il protocollo combina cinque famiglie di test — rilassamento, persistenza, isteresi, divergenza locale e change-point — sotto controlli matched-norm, seed fissati, misure indipendenti di capacità e replica fra prompt e modelli. L'obiettivo è determinare non soltanto se uno steering modifichi il comportamento, ma come l'effetto evolva, quanto persista, se dipenda dal percorso e se presenti soglie riproducibili.

L'economia cognitiva artificiale costituisce una seconda ipotesi, separata: un regolatore latente a bassa dimensionalità è informativo soltanto se comprime la descrizione di comportamenti eterogenei meglio di modelli locali, dopo penalizzazione per complessità e validazione fuori campione.

Allo stato attuale il paper è quindi un programma di ricerca falsificabile, non la presentazione di un nuovo risultato empirico. Se i test proposti non mostrassero persistenza, isteresi, divergenza o change-point oltre i controlli, il contributo si ridurrebbe a una sistematizzazione dello steering esistente. Se invece tali firme dinamiche risultassero robuste e replicabili, il protocollo fornirebbe una caratterizzazione empirica distinta della dinamica comportamentale post-intervento. Il passo successivo è un esperimento pilota su un singolo modello open-weight, con una direzione sperimentale, controlli matched-norm e le tre condizioni di persistenza della Sezione 7, che trasformerebbe il documento in un lavoro con un primo risultato empirico.

Bibliografia

- Bigelow, E., Wurgaft, D., Wang, Y., Goodman, N., Ullman, T., Tanaka, H. and Lubana, E.S. (2026) 'Belief Dynamics Reveal the Dual Nature of In-Context Learning and Activation Steering', Proceedings of the 43rd International Conference on Machine Learning, PMLR, 306, pp. 8248–8276.
- Cao, Q., Lou, J., Liu, M., Feng, W., Li, D., Ng, S.-K. and Luu, A.T. (2026) 'Activation Steering Induces Emergent Misalignment: A More Comprehensive Evaluation', arXiv:2606.08682.
- Cover, T.M. and Thomas, J.A. (2006) Elements of Information Theory. 2nd edn. Hoboken: Wiley.
- Gödel, K. (1931) 'Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I', Monatshefte für Mathematik und Physik, 38, pp. 173–198.
- Ismail, A.H., Kuang, A., Akinkugbe, A., Zhu, K. and O'Brien, S. (2026) 'CBMAS: Cognitive Behavioral Modeling via Activation Steering', arXiv:2601.06109.



- Kantz, H. and Schreiber, T. (2004) *Nonlinear Time Series Analysis*. 2nd edn. Cambridge: Cambridge University Press.
- Lai, J., Bao, A., Quinn, J. and Gilpin, W. (2026) ‘Fractal basins trap latent reasoning’, arXiv:2609.04963.
- Li, X., Leng, Y., Ding, R., Mo, H. and Yang, S. (2025) ‘Cognitive Activation and Chaotic Dynamics in Large Language Models: A Quasi-Lyapunov Analysis of Reasoning Mechanisms’, arXiv:2503.13530.
- Nakaishi, K., Nishikawa, Y. and Hukushima, K. (2024) ‘Critical Phase Transition in Large Language Models’, arXiv:2406.05335.
- Ott, E. (2002) *Chaos in Dynamical Systems*. 2nd edn. Cambridge: Cambridge University Press.
- Pearl, J. (2009) *Causality: Models, Reasoning, and Inference*. 2nd edn. Cambridge: Cambridge University Press.
- Scheffer, M., Bascompte, J., Brock, W.A. et al. (2009) ‘Early-warning signals for critical transitions’, *Nature*, 461, pp. 53–59.
- Skifstad, J., Yang, X.A. and Chou, G. (2026) ‘Local Linearity of LLMs Enables Activation Steering via Model-Based Linear Optimal Control’, *Proceedings of the 43rd International Conference on Machine Learning*, PMLR, 306, pp. 113885–113919.
- Soatto, S., Tabuada, P., Chaudhari, P. and Liu, T.Y. (2023) ‘Taming AI Bots: Controllability of Neural States in Large Language Models’, arXiv:2305.18449.
- Strogatz, S.H. (2015) *Nonlinear Dynamics and Chaos*. 2nd edn. Boulder: Westview Press.
- Tagliabue, V., Dung, L. and Berg, C. (2026) ‘The Pain Axis: LLMs Represent Self-Directed Harm and Act on It’, arXiv:2609.16247v2.
- Tarski, A. (1936) ‘Der Wahrheitsbegriff in den formalisierten Sprachen’, *Studia Philosophica*, 1, pp. 261–405.
- Turner, A.M., Thiergart, L., Leech, G., Udell, D., Vazquez, J.J., Mini, U. and MacDiarmid, M. (2023) ‘Steering Language Models With Activation Engineering’, arXiv preprint 2308.10248.
- Zou, A. et al. (2023) ‘Representation Engineering: A Top-Down Approach to AI Transparency’, arXiv:2310.01405.