Discovery report for Conceptual Crossroads: Mapping Paradigm-Level Uncertainty in Language Models

Research Objective

Run this idea for me:

Conceptual Crossroads: Mapping Paradigm-Level Uncertainty in Language Models

TL;DR: This idea explores whether token-level uncertainty sometimes stems from competing conceptual frameworks rather than just alternate reasoning paths. We'll test this by creating prompts with embedded paradigm conflicts and measuring how hidden state dynamics differ between path-level and paradigm-level uncertainty.

Research Question: When language models exhibit uncertainty during reasoning, how can we distinguish between uncertainty arising from multiple reasoning paths versus uncertainty stemming from competing conceptual frameworks or paradigms?

Hypothesis: Paradigm-level uncertainty will manifest as distinct activation patterns compared to path-level uncertainty—specifically, paradigm conflicts will produce more persistent, globally distributed activation signatures across multiple layers, while path uncertainty will be more localized and transient.

Experiment Plan: Create a curated dataset of reasoning prompts that induce either (a) multiple valid reasoning paths to the same conclusion, or (b) genuine paradigm conflicts (e.g., quantum vs. classical physics explanations). Use the activation intervention techniques from Zur et al. to map uncertainty patterns across both types. Apply Sparse Autoencoders (Galichin et al.) to identify whether paradigm conflicts activate different reasoning features than path uncertainty. Compare the persistence and distribution of uncertainty signatures across model layers and time steps. Expected outcome: Paradigm conflicts will show higher activation of "uncertainty" and "reflection" features that persist across more tokens and layers than path uncertainty.

References: ['Zur, A., Geiger, A., Lubana, E., & Bigelow, E.J. (2025). Are language models aware of the road not taken? Token-level uncertainty and hidden state dynamics.', 'Galichin, A.V., et al. (2025). I Have Covered All the Bases Here: Interpreting Reasoning Features in Large Language Models via Sparse Autoencoders. arXiv.org.']

Summary of Discoveries

Discovery 1: Dissociating Path Multiplicity and Paradigm Conflict in Hidden-State Dynamics

This report advances a mechanistic dissociation between two sources of language model uncertainty: branching among multiple valid reasoning paths versus conflict between incompatible conceptual frameworks. It synthesizes recent hidden-state analyses, intervention studies, and distribution/persistence metrics to propose testable signatures—transient, layer-localized, linearly steerable dynamics for path multiplicity versus persistent, globally distributed, and often non-linear dynamics for paradigm conflict.

Discovery 2: Confound-Controlled Induction and Causal Verification of Paradigm Conflicts

This discovery shows how to build a confound-controlled dataset and causal verification workflow that disentangles two sources of uncertainty in language models: multiple reasoning paths versus competing conceptual paradigms. It pairs balanced prompt generation and propensity score matching with behavioral validation and activation-level causal mediation to test whether paradigm conflicts induce distinct, persistent internal signatures compared to path-level uncertainty.

Discovery 3: Multivariate Activation Signatures and Supervised Classification of Uncertainty States

This discovery provides a mechanistic toolkit that distinguishes path multiplicity from paradigm conflict in language models by combining activation-based diagnostics of nonlinearity, temporal and cross-layer persistence, spatial distribution, and branching structure into supervised classifiers. Multi-view representations derived from complementary pooling schemes are co-regularized and integrated with structured penalties, enabling both diagnostic discrimination and targeted activation interventions while acknowledging limits in causal alignment and generalization.

Discovery 4: Structural Determinants and Causal Perturbation of Uncertainty Dynamics

This report proposes a structural and causal framework for distinguishing "path-level" from "paradigm-level" uncertainty in language models by linking uncertainty dynamics to network modularity, effective dimensionality, and activation topology. It integrates dose–response steering, temporal persistence, and topological markers to predict when uncertainty will be localized and linear versus globally persistent and non-linear, and outlines causal tests that modulate modularity directly.

Dissociating Path Multiplicity and Paradigm Conflict in Hidden-State Dynamics

Summary

This report advances a mechanistic dissociation between two sources of language model uncertainty: branching among multiple valid reasoning paths versus conflict between incompatible conceptual frameworks. It synthesizes recent hidden-state analyses, intervention studies, and distribution/persistence metrics to propose testable signatures—transient, layer-localized, linearly steerable dynamics for path multiplicity versus persistent, globally distributed, and often non-linear dynamics for paradigm conflict.

Background

Uncertainty quantification in language models has largely relied on output-level proxies such as token entropy, confidence, or response consistency, yet these signals are often miscalibrated and can conflate heterogeneous causes of uncertainty. Concurrently, mechanistic work has revealed that token-level uncertainty is coupled to internal dynamics, that alternative outcomes are internally represented prior to a commitment point, and that targeted activation interventions can steer model behavior along specific representational directions. These advances open the door to moving beyond surface cues to identify regime-level differences in how models encode and resolve uncertainty, with implications for diagnosis, control, and reliability.

Results & Discussion

The central discovery is that path multiplicity and paradigm conflict are distinct uncertainty regimes with different hidden-state footprints and controllability profiles. Evidence that token-level uncertainty couples to internal dynamics comes from Forking Paths analyses, where per-token outcome distributions and change-points reveal that hidden activations predict future outcome distributions and that steerability peaks before “commitment,” especially in middle layers, consistent with internal representations of alternative paths during generation [r0, zur2511]. A complementary belief-dynamics framework unifies in-context learning and activation steering: log posterior odds decompose additively into a steering magnitude term, a context-length term γN^(1−α), and an offset, yielding a phase boundary N*(m) between competing concepts; responses are often linear within a subspace and layer-localized, but can saturate or fail in some models, indicating nonlinearity or distributed encoding for some concepts [r0, bigelow2025, steeringUnknownyearbeliefdynamicsunify]. Mechanistic studies suggest superposition of parallel traces and layer-wise subtask decomposition, motivating the prediction that multiple-path uncertainty manifests as low-amplitude concurrent trajectories that are transient and localized, whereas paradigm conflicts recruit segregated high-activation subspaces and oscillatory layer dynamics that are persistent and distributed [r0, li2025]. At the behavioral level, output-level uncertainty methods frequently disagree with factuality and calibration, and deeper inference can worsen overconfidence, underscoring the need for mechanistic diagnostics beyond surface probabilities or self-reports [r0, shorinwa2025, mei2025].

This dissociation can be operationalized with a standardized measurement and intervention pipeline. First, curate prompts that selectively induce either multiple valid reasoning paths to the same conclusion or genuine paradigm conflicts (e.g., classical vs quantum explanations). Forking Paths with change-point detection provides ground truth for forking tokens, while layer-wise probes and LogitLens divergence curves localize early branching versus persistent cross-layer conflict [r0, zur2511]. The belief-dynamics boundary N*(m) supplies a quantitative phase criterion; its movement under in-context learning or steering differentiates path resolution (shifts achieved via layer-localized control) from entrenched conceptual conflict (requiring multi-layer, non-linear interventions) [r0, bigelow2025, steeringUnknownyearbeliefdynamicsunify]. Activation steering follows established dose–response protocols: compute contrastive per-layer directions, intervene h(l)t ← h(l)t

+ λv, and sweep λ while fitting linear versus sigmoid or 4-parameter log-logistic curves; model selection via AIC/WAIC, monotonicity checks, and residuals adjudicate linear versus non-linear control, with controls including random directions, orthogonal vectors, and module ablations [r4, huan2025, spinu2021, mathar2022, davis2025]. The prediction is that path-level prompts will exhibit near-linear, layer-localized dose–responses with early steerability peaks, whereas paradigm-level prompts will show steeper, switch-like non-linearity and saturation, consistent with harder-to-steer, distributed representations; notably, prior steering work has not stratified linearity by these regimes, leaving this testable difference unmeasured to date [r0, r4, zur2511].

Quantifying persistence and spread of uncertainty features strengthens the mechanistic separation. Token-time persistence can be summarized by lag-1 autocorrelation (TA- 1), multi-lag ACF, ARFIMA's fractional differencing parameter d (Hurst H = d + 0.5), detrended fluctuation analysis exponents, and 1/f spectral slopes; these capture long-range temporal correlations in activation trajectories [r6, linkenkaer-hansen2001, shinn2023]. Cross-layer persistence and agreement can be assessed with Lin's concordance, intraclass correlation, and multivariate fingerprinting, while state-space stability is indexed by recurrence-network dwell metrics (degree, clustering) and velocity proxies derived from successive distances [r6, shinn2023, varley2022]. Topological "neural persistence" complements these by summarizing structural complexity growth across layers via persistent homology of weight filtrations [r6, rieck2018]. The hypothesis predicts higher ACF/DFA/Hurst, stronger cross-layer concordance/ICC, longer dwell times, and broader structural persistence for paradigm conflicts than for path multiplicity, which should present as short-lived, layer-localized divergences with lower temporal and cross-layer persistence [r0, zur2511].

Spatial distribution at single timesteps can be characterized using attention-distribution Gini coefficients (concentration), effective dimensionality (coordination), mutual information and total correlation among heads/neurons (dependence), modularity Q of interaction graphs (clustering), ablation-normalized impact (functional dependence), and heavy-tail indices (e.g., Hill tail α) on spectral statistics [r7, liu2509, huang2508, hod2021]. In parallel, response-graph spectra and cluster entropy over outputs provide coarse signatures of branching versus overlapping conceptual groups, and sparse autoencoders can disentangle polysemantic features that underwrite paradigm entanglement for selective interventions [r0, shorinwa2025]. Under the proposed hypothesis, paradigm conflicts should exhibit broader, coordinated recruitment across heads and layers (lower within-layer concentration, higher cross-component dependence, and wider footprint across layers), whereas path multiplicity should concentrate uncertainty signals within a smaller subset of heads and layers and remain more amenable to linear, layer-localized control [r0, r7, huang2508].

The significance of this dissociation is twofold: it offers a mechanistic diagnostic to separate superficially similar uncertainty behaviors and it enables regime-appropriate controls. Output-level uncertainty methods are variably calibrated and can worsen overconfidence with deeper inference, so hidden-state diagnostics are necessary to avoid conflating causes [r0, shorinwa2025, mei2025]. Several caveats inform rigorous application: autocorrelation inflates degrees of freedom and time-varying correlations can be spurious without permutation nulls; MI estimation can be fragile; and steering can saturate or misalign at large λ, sometimes requiring multi-directional steering or position-dependent schedules [r6, r7, hutchison2013, huan2025, davis2025]. Notably, no prior work reports definitive neural activation signatures separating paradigm conflicts from path uncertainty, emphasizing the value of curated stimuli that manipulate each factor independently and composite metrics that integrate persistence, spread, and controllability [r0]. Finally, contextual control can arbitrate between "intuitive" and "deliberative" modes and cognitive modularity suggests distinct internal signatures when different trackers are engaged, further motivating a paradigm/path distinction and its practical use for improving reliability and steerability in complex reasoning [r0, lampinen2024, mahowald2024].

Trajectory Sources

Trajectory r0: Overview and motivation. Uncertainty in LLM reasoning is heterogeneous: variability can reflect branching choices among valid inference chains or deeper conflicts between incompatible conceptual frames, with important implications for diagnosis, control, and reliability of model outputs (path multip...

Trajectory r4: The literature provides step-by-step activation-steering and dose–response methodologies to test linear versus non-linear control, but it does not operationalize or measure differences between path- and paradigm-level uncertainty, so the hypothesis is only partially supported (huan2025 pa...

Trajectory r6: Yes—the neuroscience, signal-processing, and time-series literatures define mature metrics for temporal and cross-“depth” (layerwise/structural) persistence, including autocorrelation/long-memory measures for sequences, concordance/ICC/fingerprinting for cross-condition consistency, and persistent-h...

Trajectory r7: Yes—multiple studies explicitly propose and apply quantitative metrics such as Gini coefficients, entropy-based measures, mutual information/total correlation, modularity, effective dimensionality, and spectral tail indices to quantify how activations or information are spatially distributed across ...

Confound-Controlled Induction and Causal Verification of Paradigm Conflicts

Summary

This discovery shows how to build a confound-controlled dataset and causal verification workflow that disentangles two sources of uncertainty in language models: multiple reasoning paths versus competing conceptual paradigms. It pairs balanced prompt generation and propensity score matching with behavioral validation and activation-level causal mediation to test whether paradigm conflicts induce distinct, persistent internal signatures compared to path-level uncertainty.

Background

Language models can be uncertain because they are exploring several valid chains of reasoning or because the task invokes incompatible conceptual frameworks. Distinguishing these regimes requires experimental control: prompts must reliably induce either path multiplicity or paradigm conflict while holding linguistic complexity and intrinsic difficulty constant, and internal activations must be linked to behavior with causal evidence. Recent advances in prompt design for multi-path reasoning, persona- and evidence-driven paradigm opposition, internal activation analysis, and mediation-based causal inference together make such a study tractable at scale.

Results & Discussion

The core contribution is an end-to-end protocol that induces and verifies paradigm conflicts while controlling for confounds, enabling causal tests that connect internal activation patterns to behavior. First, two divergent prompt-generation pipelines target distinct uncertainty regimes. For path multiplicity, established methods such as self-consistency, Tree/Graph-of-Thoughts, decomposed prompting, and program-structured reasoning are used to elicit multiple diverse yet valid chains that converge on the same answer; these frameworks yield robust gains over single-path prompting, for example ≈11%/3%/6% improvements on math, commonsense, and multi-hop tasks with self-consistency, and a 74% vs 4% win on Game of 24 with Tree-of-Thoughts, confirming that prompts and decoding can intentionally produce multi-path reasoning [r1, chen2310, sahoo2024, chen2025, choi2025, vatsal2407, wei2022]. For paradigm conflict, persona and role conditioning, multi-agent dialectic setups, counterfactual/comparative framing, and contradiction via in-context learning or retrieval are used to instantiate competing stances; the literature provides actionable scaffolds but emphasizes fragility, order sensitivity, and the need for methodological hygiene to ensure that observed conflicts reflect genuine conceptual differences rather than prompt artifacts [r2, peng2505, simons2506, kumar2025]. Together these pipelines furnish large pools of candidate prompts designed to differentially induce path-level versus paradigm-level uncertainty [r49, kepel2407].

Second, the discovery introduces a confound-control layer that balances linguistic complexity and intrinsic reasoning difficulty across the two prompt families before any internal analyses. Linguistic covariates—drawn from validated toolkits such as L2SCA and Coh-Metrix—include mean sentence length, T-unit and clause ratios, dependency metrics, lexical density and sophistication, and information-theoretic surprisal/entropy reduction, all with established computational implementations [r46, lu2017, lu2019, hale2016]. Intrinsic difficulty covariates are drawn from domain-agnostic metrics such as reasoning depth/steps, branching factor, search effort, working-memory load, automaton/memory class, and compositional/retrieval hops; benchmark generators like Multi-LogiEval and LogicBench validate that performance degrades with depth and rule composition, and task facets for matching the two regimes include minimal branching factor or number of valid derivations (path multiplicity) versus counts of negations/defeasible exceptions and causal level (paradigm conflict) [r39, helie2022, isaac2014, hernandezorallo2014, patel2024, parmar2024, bellodi2025, sun2025, mondorf2404]. Prompts are then balanced via propensity score matching using logistic regression over this covariate set (nearest neighbor or optimal matching), producing pairs that differ primarily in conceptual induction rather than language or difficulty; controls that target contamination (e.g., surprisal-based regeneration, symbol remapping) are acknowledged but distinguished from intrinsic difficulty calibration [r49, frincu2023, yax2024, ma2505].

Third, the protocol behaviorally verifies that each matched prompt reliably elicits the intended conceptual state. A rubric-driven LLM-as-judge with mandatory self-explanation is used to score conflict detection, stance articulation, and resolution quality; such evaluators achieve 78% agreement with human experts in Socratic, multi-turn settings and outperform keyword heuristics, with further gains from committee aggregation and iterative calibration [r72, multitrun2025, cao2025, zhang2025b]. Scores from multiple judges are aggregated with a Bayesian Dawid–Skene model that estimates judge reliabilities and reduces bias relative to mean or majority voting, with open-source implementations available and mode estimators performing best on hard evaluation tasks [r81, gao2024b, yao2024]. This behavioral layer—augmented by prompt minimalism, multi-response sampling, and transcript preservation—reduces variance and helps ensure that measured effects are substantive and not artifacts of prompt wording or sampling order [r2, peng2505, kumar2025].

Finally, the workflow links behavior to internal mechanisms and tests causality. Internal activation analysis validates concept induction using linear probes and Concept Activation Vectors to measure alignment of hidden states with predefined concepts, and causal interventions (activation patching, ablations, gating) test the internal impact of prompts independent of outputs [r10, lee2025, davies2408, dalal2024]. On top of this, a causal mediation analysis (CMA) framework treats prompt type as the treatment, activation-derived metrics as mediators, and rubric-based scores as outcomes; mediator families explicitly include persistence (temporal/layerwise durability of "uncertainty/reflection" features) and modularity/distribution (community structure and spread across heads/layers) [r71, mueller2024, rocchetti2024]. The CMA pipeline specifies a structural model x→z→y, estimates total/direct/indirect effects via a mix of observed and interventional runs, and executes do(z=z2) using input-dependent substitutions or ablations; uncertainty is addressed with path stability scores, Bayesian model averaging, and cross-fitted semiparametric estimators, alongside sensitivity analyses for on-manifoldness and SUTVA/positivity assumptions [r71, mueller2024, liUnknownyeardecodingcausalstructure, liu2025a]. This enables a direct test of the central hypothesis—that paradigm conflicts produce more persistent and globally distributed activation signatures than path-level uncertainty—by quantifying whether persistence/modularity mediate a significant share of the prompt effect on behavior; proposed tests expect mid-to-late-layer mediators to account for substantial indirect effects and to be more stable with input-dependent substitutions than mean ablations [r71]. The overall result is a confound-controlled dataset and causally grounded verification workflow that together allow precise adjudication of paradigm-level versus path-level uncertainty in language models [r49, mueller2024, rocchetti2024].

Trajectory Sources

Trajectory r1: The hypothesis is supported: the literature explicitly describes prompt and decoding techniques—most notably self-consistency, Tree/Graph-of-Thoughts, decomposed prompting, and program-structured reasoning—that elicit multiple distinct yet valid reasoning chains which are then aggregated to the same...

Trajectory r2: The hypothesis is partly supported: peer-reviewed and preprint literature provides concrete, replicable prompting methods (persona/role conditioning, multi-agent/dialogic setups, counterfactual framing, RAG-fed contradictions, and in-context contradictory exemplars), but detailed, step-by-step proto...

Trajectory r10: The existing literature confirms that internal activation analysis techniques—such as linear probing, feature activation inspection, and causal intervention—provide robust methods to verify that a prompt or in-context examples activate the intended conceptual representations independent of final out...

Trajectory r39: Partially supported: robust, domain-independent operational metrics for intrinsic reasoning complexity exist, but their application to LLM prompts as an explicit control for task difficulty is limited to parameterized benchmarks (e.g., depth/rule isolation) rather than a unified, cross-benchmark dif...

Trajectory r46: The literature confirms that specific, operational metrics exist for quantifying both linguistic complexity and intrinsic reasoning difficulty, many of which have validated software implementations or detailed computational descriptions for reimplementation (lu2017 pages 10-13,...

Trajectory r49: A protocol that generates a large pool of candidate prompts using divergent methods for path multiplicity and paradigm conflict, scores them with external r46 linguistic/difficulty metrics, and matches them via propensity score techniques while verifying intended conceptual states with a suite of f6...

Trajectory r71: Yes—recent interpretability, causal-inference, and econometrics papers collectively provide a complete, step-by-step CMA workflow that directly supports testing whether activation-derived metrics such as persistence and modularity mediate the effect of paradigm-conflict prompts on quantitative behav...

Trajectory r72: The hypothesis is provisionally supported: rubric-driven LLM-as-judge with self-explanation (chain-of-thought) and multi-evaluator aggregation shows higher alignment with human experts for complex dialogue-like behaviors than keyword heuristics and, while head-to-head evidence is scarce, appears mor...

Trajectory r81: The literature indicates that a Bayesian Dawid–Skene model, particularly when calibrated with informative priors, is best suited for committee-based LLM evaluations, producing scores that align more closely with expert human ratings than simple mean or majority voting (gao2024b p...

Multivariate Activation Signatures and Supervised Classification of Uncertainty States

Summary

This discovery provides a mechanistic toolkit that distinguishes path multiplicity from paradigm conflict in language models by combining activation-based diagnostics of nonlinearity, temporal and cross-layer persistence, spatial distribution, and branching structure into supervised classifiers. Multi-view representations derived from complementary pooling schemes are co-regularized and integrated with structured penalties, enabling both diagnostic discrimination and targeted activation interventions while acknowledging limits in causal alignment and generalization.

Background

Uncertainty in language model reasoning reflects heterogeneous causes—some tokens encode branching among multiple valid inference chains, while others reflect competition between incompatible conceptual frames—yet common token-level proxies (e.g., entropy) blur these distinctions and can be miscalibrated with respect to claim-level correctness. Recent advances in mechanistic interpretability show that hidden-state dynamics predict outcome distributions, admit controlled steering, and sometimes reveal linear subspaces, suggesting that internal uncertainty states are measurable and manipulable. These developments motivate a principled pipeline that maps how uncertainty is represented in activations, tests linear versus nonlinear controllability, and aggregates heterogeneous metrics into supervised diagnostics that differentiate path-level ambiguity from paradigm-level conflict.

Results & Discussion

The core contribution is a validated, multivariate activation-signature toolkit that separates path-level from paradigm-level uncertainty by design. It leverages recent evidence that hidden activations encode the distribution over future outcomes and exhibit “forking” dynamics, with predictability and steerability peaking before commitment in mid-layers, thereby providing groundable labels of path branching via change-point detection and probes (Forking Paths with CPD, LogitLens, linear probes) [r0, zur2511]. A complementary belief-dynamics framework relates additive activation steering to context length, yielding a phase boundary N*(m) between competing concepts; linear subspace control is often layer-localized but saturates or fails in some models, consistent with distributed or nonlinear encoding expected under paradigm conflict [r0, bigelow2025, steeringUnknownyearbeliefdynamicsunify]. Mechanistic accounts of superposition and layer-wise subtask decomposition supply candidate signatures: low-amplitude concurrent trajectories for multiple-path uncertainty and segregated, oscillatory, higher-activation subspaces for paradigm conflict, measurable with probes and logit-lens analyses [r0, li2025]. Because output-level uncertainty tools are variably aligned with factuality and can overstate confidence, the pipeline elevates mechanistic diagnostics over surface probabilities while using calibration and introspective checks as outcome-level validations [r0, shorinwa2025, mei2025]. Together, these findings motivate operational tests in which paradigm conflict is hypothesized to manifest as persistent, distributed, and less linearly steerable activation patterns, whereas path multiplicity produces transient, layer-localized, and nearly linear divergences [r0, shorinwa2025, zur2511].

Nonlinearity and steering controllability form the first diagnostic axis. Dose–response methodologies intervene as h(l)t ← h(l)t + λv at targeted layers while sweeping λ, with controls and curve-fitting to adjudicate linear versus nonlinear control (e.g., monotone vs plateauing/sigmoidal responses) [r4, huan2025, spinu2021]. Axbench-style evaluations show that simple closed-form concept vectors—difference-in-means and linear-probe-derived directions (including ReFT variants)—consistently outperform PCA or sparse-autoencoder features for causal steering of high-level concepts, providing robust, reproducible baselines for the intervention component of the toolkit [r44, wu2025a, zhao2025]. This supports a pragmatic division of labor: SAEs are prioritized for feature discovery and disentanglement of polysemantic representations linked to paradigm entanglement, while steering relies on difference-in-means or probe directions that deliver stronger causal efficacy [r0, r44, shorinwa2025, wu2025a]. Notably, although graded steering and curve-fitting protocols are mature, existing studies have not explicitly stratified linearity by path versus paradigm categories, underscoring the need for the present operationalization to quantify regime-specific dose–response signatures [r4, huan2025].

Persistence, spatial distribution, and branching structure supply orthogonal axes that together differentiate uncertainty regimes. Temporal persistence is quantified via autocorrelation functions (including lag-1 statistics), long-range dependence estimates (ARFIMA d/Hurst exponent), detrended fluctuation analysis scaling, and 1/f spectral slopes; cross-layer persistence is summarized with Lin's concordance, intraclass correlation, and multivariate fingerprinting across layers or model variants [r6, shinn2023, linkenkaerhansen2001]. State-space stability and dwell-time are captured by recurrence-network measures (degree, clustering, velocity proxies), complementing topological summaries of structural stability such as neural persistence [r6, varley2022, rieck2018]. Spatial distribution at single time steps is characterized by inequality and dependence metrics over neurons/heads/layers: Gini coefficients, mutual information and total correlation, modularity, effective dimensionality, and spectral tail indices on weight/representation spectra; these measures quantify concentration versus distribution and coordination among components [r7, liu2509, huang2508]. Branching versus blending is assessed with response-graph spectra, modularity, community counts, Laplacian spectral gaps, and eigenvector localization, where discrete branches exhibit higher modularity, fewer well-defined communities, and larger spectral gaps than diffuse mixtures—an expected hallmark of path multiplicity versus paradigm blending [r8, kolchinsky2015, chakraborti2024, volchenkov2025, laskaris2020]. Pooling choices materially affect these summaries: max pooling over selected intermediate/final layers improves robustness and preserves salient signals relative to mean pooling, and learnable or order-aware aggregation can be benchmarked as a secondary option [r16, behrendt2505, gulko2025, bashier2023]. The toolkit therefore computes multiple pooled "views" (e.g., token-local, layer-aggregated, state-space) and treats them as complementary inputs for supervised classification; added demonstrations and EU/AU decomposition, together with movement of the belief-dynamics phase boundary N*(m), further test reducibility (path) versus persistence (paradigm) of uncertainty under evidence [r0, shorinwa2025, wang2025, bigelow2025].

Finally, the pipeline integrates heterogeneous metrics into a supervised classifier with theory-guided feature grouping and multi-view co-regularization. Composite time-series/network feature vectors are a validated strategy for state prediction, and are here extended to LLM activations by combining nonlinearity indices (dose–response fits), persistence metrics, spatial-distribution statistics, and graph-branching signatures [r13, raubitzek2021, clough2025]. To promote parsimony, stability, and interpretability, the classifier uses structured penalties such as group LASSO or sparse group LASSO on feature families (e.g., "persistence," "distribution," "branching," "nonlinearity"), which outperform unstructured penalties when predictors are meaningfully grouped and correlated [r48, ajana2019, munch2021, emmert-streib2019]. Multi-view integration is achieved with cooperative learning that co-regularizes agreement across pooled views and includes a standardized group-lasso block penalty for view elimination, improving over naive concatenation or late fusion when views share latent signal [r55, ding2022, xu2013, loon2020]. Targeted interventions are then selected by aligning classifier-discriminative features with causally effective steering directions (difference-in-means/probe vectors) to nudge the model out of path ambiguity or to deploy broader, multi-layer interventions for paradigm conflicts; however, the field lacks definitive neural signatures separating these regimes and steering can saturate or behave nonlinearly in some models, so causal validation, calibration checks, and generalization tests remain essential [r0, r4, r44, zur2511, steeringUnknownyearbeliefdynamicsunify, mei2025, wu2025a, huan2025].

Trajectory Sources

Trajectory r0: Overview and motivation. Uncertainty in LLM reasoning is heterogeneous: variability can reflect branching choices among valid inference chains or deeper conflicts between incompatible conceptual frames, with important implications for diagnosis, control, and reliability of model outputs (path multip...

Trajectory r4: The literature provides step-by-step activation-steering and dose–response methodologies to test linear versus non-linear control, but it does not operationalize or measure differences between path- and paradigm-level uncertainty, so the hypothesis is only partially supported (huan2025 pa...

Trajectory r6: Yes—the neuroscience, signal-processing, and time-series literatures define mature metrics for temporal and cross-"depth" (layerwise/structural) persistence, including autocorrelation/long-memory measures for sequences, concordance/ICC/fingerprinting for cross-condition consistency, and persistent-h...

Trajectory r7: Yes—multiple studies explicitly propose and apply quantitative metrics such as Gini coefficients, entropy-based measures, mutual information/total correlation, modularity, effective dimensionality, and spectral tail indices to quantify how activations or information are spatially distributed across ...

Trajectory r8: Graph-based analyses of model outputs or internal states do reveal clear structural differences between a few discrete branching choices and a diffuse mixture of concepts.

Trajectory r13: The literature supports that frameworks exist for integrating multiple, conceptually distinct time-series metrics into composite feature vectors for supervised state classification, although explicit examples using network-specific metrics (like Gini coefficient or modularity) remain less common. (r...

Trajectory r16: The evidence indicates that both the aggregation function and the choice of layers significantly affect computed persistence and distribution metrics, with max pooling over select intermediate layers consistently yielding more robust summarizations than mean pooling. (behrendt2505maxpoolbertenhancin...

Trajectory r44: Recent head-to-head benchmarking studies, notably Axbench, show that simple closed-form approaches—particularly difference-in-means and linear probe weight methods (including ReFT variants)—consistently yield the most causally effective and robust steering of high-level concepts in LLMs (wu2025axben...

Trajectory r48: The surveyed literature supports that structured penalties—such as group LASSO and sparse group LASSO—yield models with higher predictive performance, more stable variable selection, and improved interpretability compared to unstructured methods in settings with meaningful, highly correlated predict...

Trajectory r55: Choose cooperative learning with an explicit agreement (co-regularization) penalty, augmented with a standardized group-lasso block penalty for view elimination, as the primary integration method; this supports the hypothesis that consensus/complementarity–aware methods outperform concatenation or l...

Structural Determinants and Causal Perturbation of Uncertainty Dynamics

Summary

This report proposes a structural and causal framework for distinguishing "path-level" from "paradigm-level" uncertainty in language models by linking uncertainty dynamics to network modularity, effective dimensionality, and activation topology. It integrates dose–response steering, temporal persistence, and topological markers to predict when uncertainty will be localized and linear versus globally persistent and non-linear, and outlines causal tests that modulate modularity directly.

Background

Language models often display token-level uncertainty arising from multiple plausible continuations, but it remains unclear when this uncertainty reflects mere competition among alternative reasoning paths versus deeper conflicts between incompatible conceptual frameworks. Across complex systems, structural properties such as modularity and dimensionality predict whether perturbations remain localized or cascade non-linearly, suggesting that analogous principles may govern uncertainty dynamics in large language models. Recent advances in activation steering, time-series analysis, topological data analysis, and continuous-time mediation provide a toolkit to measure the persistence, distribution, and causality of hidden-state dynamics, enabling an operational distinction between path- and paradigm-level uncertainty.

Results & Discussion

The central claim is that structural determinants—modularity, effective dimensionality, and integration—forecast whether perturbations to hidden states elicit localized, approximately linear responses or precipitate non-linear cascades, and that this structure–response link can be used to separate path- from paradigm-level uncertainty in language models. Empirical work in networked systems shows that higher modularity buffers perturbations and confines them to communities, whereas more integrated architectures with connector hubs and assortativity are prone to synchronized or cascading tipping transitions [r25]. Complementary evidence from genotype→phenotype→fitness maps, neuronal pharmacology, and supervised latent-factor modeling demonstrates that low effective dimensionality and modular sparsity favor linear, predictable dose–responses, while pleiotropy, higher dimensionality, and shifting module engagement produce non-linear or nonmonotonic curves; for example, selectively sampling a distinct module (a low-dose condition) restored linear predictability and increased out-of-sample R^2 by ~0.2 in a challenging combination [r14]. These principles transfer to language models by constructing functional networks in which nodes correspond to neurons or modules and edges derive from activation co-occurrence or attention weights; standard graph metrics (modularity Q, within-module degree, participation coefficient) differentiate segregated from integrated processing states and thus provide structural context for uncertainty regimes [r36]. Together, these results motivate a testable prediction: path-level uncertainty should appear in modular, low-dimensional regimes (localized, more linear responses), whereas paradigm-level uncertainty should manifest in more integrated, higher-dimensional regimes (globally distributed, non-linear cascades) [r14, r25, r36].

The report operationalizes persistence and distribution using mature temporal, structural, and topological metrics. Temporal persistence along the token sequence can be quantified by the lag-1 autocorrelation TA-Δ1 and the full ACF; long-range dependence is captured via ARFIMA's fractional differencing parameter d and the Hurst exponent $H = d + 0.5$, with detrended fluctuation analysis and $1/f$ spectral slopes as complementary indices; practical computation accommodates gapped sequences and misalignment via segment averaging and alignment kernels (e.g., DTW) [r6]. Cross-layer or cross-"depth" persistence is characterized with Lin's concordance, intraclass correlation, and multivariate fingerprinting, while state-space stability is summarized by recurrence-network measures (degree/occupancy, clustering, and a velocity proxy from successive distances) [r6]. Topology provides orthogonal markers: persistent-homology features (mean birth/death, persistence counts, and persistent entropy) separate clean versus adversarial or "elicited" versus "locked" states across models, with adversarial/locked regimes exhibiting "topological compression"—fewer, later-born, longer-lived loops and reduced persistent entropy—and PH-based classifiers achieving ROC–AUC 1.00 ± 0.00 under held-out testing and adaptive LLMail-Inject attacks, with H0 mean death and H1 count as dominant predictors [r26]. Critically, time-series persistence and topology can be linked directly: a weighted natural visibility graph→PH pipeline yields lifetime-based estimators of H that are robust to sample length, and sliding-window dynamic-TDA shows persistent entropy tracks adjusted persistence $(2 - \mathrm{HFD})$ with permutation-based significance testing, enabling joint topology–persistence profiling on activations [r62]. Under the proposed framing, paradigm conflicts should display higher token-time persistence (larger ACF/H), stronger cross-layer concordance, and topological regime shifts (e.g., compressed loop spectra) than path competition, which is expected to be more transient and layer-localized [r6, r26, r62].

Causal dose–response protocols make these distinctions testable. Existing activation-steering pipelines construct contrastive prompt pairs, compute per-layer difference vectors or PCA directions, intervene as $\mathrm{h(l)t} \leftarrow \mathrm{h(l)t} + \lambda \mathrm{v}$ at selected layers (localized by causal tests), and sweep λ while measuring behavioral metrics such as honesty rate P(truth) and liar score; dose–response curves are then fit with linear, polynomial, or sigmoid families, with formal model selection via AIC/WAIC and controls including random/orthogonal directions and module ablations [r4]. These studies report both monotonic increases (e.g., honesty rising with λ) and non-linear plateaus at higher strengths, and they recommend multi-directional steering if a single vector poorly aligns with the target manifold; discrete ablations provide coarser, graded control that complements continuous steering [r4]. Within this framework, the specific, testable prediction is that prompts inducing paradigm-level uncertainty will yield steeper, more switch-like 4-parameter log-logistic fits (larger Hill slopes) selected by WAIC, whereas path-level prompts will be better captured by linear fits or sigmoids with smaller slopes; further, an exponentially decaying multi-position schedule should increase linear-fit R^2 and reduce saturation for path-level prompts [r4]. These dose–response signatures tie back to structure: higher effective dimensionality and integration predict non-linear control (paradigm conflicts), while modularity and sparsity support linear regime behavior (path competition) [r14, r25].

Beyond measurement, the report outlines causal structural manipulation of uncertainty dynamics by directly steering modularity. Differentiable modularity proxies and spectral relaxations enable optimization of an activation direction d at a chosen layer such that adding $\alpha \cdot \mathrm{d}$ systematically increases or decreases modularity Q; practical workflows define a modularity loss on an adjacency built from activations or gradient-based similarities, optimize $\mathrm{J(d)} = \pm\lambda \cdot \mathrm{ModularityProxy(h\ (x)}{+}\alpha\mathrm{d)}$ with norm constraints to preserve task performance, and validate with pre/post clusterability and effective-circuit metrics [r76]. This provides a direct test of the hypothesis that increasing modularity suppresses globally persistent, non-linear uncertainty signatures, shifting the system toward localized, linear responses indicative of path-level uncertainty [r76]. Geometry and topology offer complementary predictions about intervention responsiveness: layers showing larger increases in H0 mean death and decreases in H1 loop counts under attack tend to be more recoverable by a single linear steering vector, and lower persistent entropy is hypothesized to predict more linear steering behavior, though direct, causal demonstrations of topological manipulation via activation patching remain an open challenge [r26, r69]. Together, modularity steering and PH diagnostics close the loop between structure, topology, and causal control.

Finally, continuous-time structural equation modeling supplies a principled mediation analysis to distinguish direct from indirect propagation of uncertainty across layers or modules. With a drift matrix A describing activation dynamics, the total effect of a unit pulse on source i at time t on target j at lag τ is the (j,i) entry of e^{A τ}; the direct effect with mediator pathways removed is obtained by exponentiating a modified drift A(D[−k]), and the indirect effect is defined as the difference, preserving total = direct + indirect across τ [r95]. Because these effects depend on the eigenstructure of A, mediated influences can vary non-monotonically with time, making the lag profile informative for discriminating localized path competition from distributed paradigm conflict [r95]. Inference should guard against inflated degrees of freedom from autocorrelation by using segment-wise handling, permutation controls, and multiple-comparisons correction, as recommended in high-dimensional neuroimaging analyses [r6]. Notably, while activation steering and dose–response tooling are mature, the literature does not yet operationalize "path- versus paradigm-level uncertainty"; the present framework fills this gap by combining structural predictors (modularity/dimensionality), temporal and topological persistence metrics, causal steering (including modularity control), and dynamic mediation into a cohesive experimental program for mapping paradigm-level uncertainty in language models [r4, r6, r14, r25, r26, r36, r76, r95].

Trajectory Sources

Trajectory r4: The literature provides step-by-step activation-steering and dose–response methodologies to test linear versus non-linear control, but it does not operationalize or measure differences between path- and paradigm-level uncertainty, so the hypothesis is only partially supported (huan2025 pa...

Trajectory r6: Yes—the neuroscience, signal-processing, and time-series literatures define mature metrics for temporal and cross-"depth" (layerwise/structural) persistence, including autocorrelation/long-memory measures for sequences, concordance/ICC/fingerprinting for cross-condition consistency, and persistent-h...

Trajectory r14: Yes—across yeast drug-resistance maps, in vitro neuronal networks, and chemical structure–activity modeling, linear versus nonlinear dose–response behavior is empirically tied to internal structural metrics, including low effective dimensionality and modular/environment-specific sparsity (favoring l...

Trajectory r25: Empirical and theoretical evidence confirms that key graph-structural metrics such as modularity, centrality, and the distribution of connector hubs are predictive of how networks respond dynamically to perturbations, with highly modular systems exhibiting more contained, linear responses while inte...

Trajectory r26: Geometric and topological summaries of activation manifolds (e.g., persistent-homology barcodes, separability/dispersion metrics) reliably predict model responses under adversarial and behavioral interventions and align with linear separability, providing indirect support that they predict when line...

Trajectory r36: The evidence confirms that LLM layers and whole models have been modeled as functional networks by defining nodes—whether as individual neurons, neuron groups, or modules—and edges derived from activation correlations, attention weights, or feature co-occurrences, with standard network metrics (such...

Trajectory r62: Yes—there are concrete theoretical and methodological papers that propose adaptable experimental designs to test correlations between TDA-derived topological summaries (e.g., persistent entropy) and time-series persistence metrics (e.g., Hurst exponent) on neural activation time series.

Trajectory r69: The current literature offers indirect, loss-based techniques (e.g., gradient-steered topological regularizers and differentiable topology layers) that influence activation topology, but it lacks explicit demonstrations using activation patching or direct gradient-based steering to causally modify h...

Trajectory r76: The hypothesis is supported: the literature provides differentiable modularity proxies and optimization machinery that can be repurposed to learn an activation steering vector d whose addition $\alpha \cdot$ d to a layer reliably increases or decreases modularity Q. (golechha2025 pages 3-4, gole...

Trajectory r95: Yes—the CT-SEM literature gives a concrete, implementable procedure: compute total effects as the (j,i) entry of e^{A τ}, compute direct effects by zeroing mediator paths in A and exponentiating, and define the indirect effect as their difference across any lag τ. (ryan2022 pages 29-3...