PHASE GEOMETRY OF REASONING: PREDICTING AND CONTROLLING EVIDENCE INTEGRATION IN LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models often produce different probabilistic judgments when equivalent evidence is reordered, reframed, or combined, yet the latent thought dynamics responsible for these changes remain poorly understood. We operationalize a thought as an atomic, belief-changing state transition induced by processing one evidence item, and represent reasoning as a trajectory through a latent state space. We hypothesize that these trajectories may follow complex-valued, noncommutative dynamics in which phase-dependent interference determines how successive thoughts combine. To test this hypothesis, we introduce Complex Probabilistic Interventional Reasoning (CPIR), a controlled framework that traces model judgments across reasoning sequences with known priors and likelihood ratios. CPIR intervenes on individual thought updates through evidence masking, polarity reversal, strength modification, order permutation, paraphrasing, and compositionally held-out combinations of these operations. We compare complex-valued state-space models with normative Bayesian inference, logistic interaction models, nonlinear real-valued recurrent models, phase ablations, and parameter-matched controls. On five frozen synthetic environments with planted complex-coherent dynamics, CPIR identifies the correct compositional family and predicts held-out domains, paraphrases, and longer sequences, reducing joint-holdout logit RMSE from 0.072 to 0.023 relative to the strongest real-valued baseline. When transferred to Qwen3-14B and DeepSeek-V3.2 judgments, the same prespecified procedure discovers target-specific real and heterogeneous predictive regimes, demonstrating selective model assignment across targets. Extending CPIR to sequential LLM evaluation, repeatability and competence controls reveal that the observed cross-order reversal rate (11.1%) is matched by same-order variation (10.0%) for the competence-qualified judge, motivating checklist elicitation that reduces reversals to 0.7%. At equal inference cost, averaging opposite criterion orders improves Brier score from 0.156 to 0.146 over averaging two same-order evaluations, with the gain concentrated in probabilistic calibration. Together, these results show that CPIR can recover structured thought composition, discriminate among compositional regimes, and improve the stability and calibration of sequential LLM evaluation.

1 INTRODUCTION

Large language models make consequential judgments by processing information sequentially. They integrate evidence, revise probabilities, compare candidate responses, and use those evaluations to guide later decisions. These judgments are often sensitive to order: permuting few-shot demonstrations can move performance from near chance to near state of the art (Zhao et al., 2021; Lu et al., 2022), and swapping the presentation order in an LLM-as-a-judge prompt can change its preference (Zheng et al., 2023; Wang et al., 2024). If two meaningful updates produce different judgments when composed as $A \rightarrow B$ rather than $B \rightarrow A$, the observation creates an opportunity to identify a noncommutative reasoning process. Realizing that opportunity requires separating structured composition from sampling variation, prompt ambiguity, evaluator competence, and flexible retro-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

spective fits. CPIR supplies this separation through intervention, held-out prediction, and matched controls.

Chain-of-thought prompting can improve task accuracy (Wei et al., 2022; Wang et al., 2023), while targeted interventions offer a more direct test of whether a reasoning step or representation supports the prediction (Vig et al., 2020; Geiger et al., 2022; Tutek et al., 2025). We therefore operationalize a *thought* as an atomic, belief-changing state transition induced by processing one controlled evidence item, rather than as a privileged transcript of hidden computation (Jacovi & Goldberg, 2020; Turpin et al., 2023). A reasoning episode becomes a sequence of manipulable interventions and observed probabilistic judgments. This definition makes thought composition experimentally testable even when free-form explanations provide an incomplete view of the model's internal algorithm.

We study whether these thought transitions admit a stable compositional law. Let p_0 be the stated prior and let $\tau_{1:k}$ denote an ordered sequence of atomic thoughts. The core complex hypothesis assigns each thought type an operator $T_{\tau_t} \in \mathbb{C}^{2\times 2}$ and predicts

$$z_0 = \left[\sqrt{p_0} \quad \sqrt{1-p_0}\right]^\top, \qquad z_t = \frac{T_{\tau_t} z_{t-1}}{\|T_{\tau_t} z_{t-1}\|_2}, \qquad \widehat{p}(H \mid \tau_{1:k}) = \left|e_1^\top z_k\right|^2. \tag{1}$$

Order effects arise when $T_b T_a \neq T_a T_b$; phase preserves directional information that can alter later transitions even when two intermediate states yield similar scalar probabilities. Related complex-valued models exploit this rotational structure as an inductive bias (Trabelsi et al., 2018), and quantum-probability models use noncommuting operations to predict order effects in human judgment (Wang et al., 2014). Our hypothesis is representational and predictive: because $T = A + iB$ has the exact real-coordinate form $\left[\begin{smallmatrix} A & -B \\ B & A \end{smallmatrix}\right]$, evidence for the complex parameterization must come from more accurate prediction of unseen interventions under matched information and capacity. This formulation tests a structured inductive bias without requiring a quantum or intrinsically complex physical interpretation of transformer activations.

To answer this question, we introduce *Complex Probabilistic Interventional Reasoning* (CPIR). CPIR constructs controlled evidential sequences with known priors and likelihood ratios, records the model's probability after composed thought updates, and generates interventions by masking evidence, reversing polarity, modifying strength, permuting order, paraphrasing surface form, and combining operations. Its central design principle is compositional holdout: a model must extrapolate to unseen combinations, longer sequences, new domains, and held-out paraphrases rather than merely interpolate among observed orders. We compare the complex state-space hypothesis with normative Bayesian inference, logistic interaction models, nonlinear real-valued recurrent models, phase and coherence ablations, parameter-matched models, and exact real-coordinate implementations. Predictive gates frozen before collection specify which outcomes would support a complex-specific claim.

The experiments deliberately separate *identifiability* from *empirical support*. Across five frozen synthetic environments generated by planted complex-coherent dynamics, the structured complex model recovers the law from textual evidence descriptions and final probabilities. It reduces mean joint-holdout logit RMSE from 0.072 for the strongest real-valued baseline to 0.023, and generalizes to held-out domains, paraphrases, and sequences of five to eight updates. This establishes the primary positive result: CPIR can identify the correct compositional family from observable judgments and extrapolate its dynamics beyond the training support. Applied to Qwen3-14B (Qwen Team, 2025) and DeepSeek-V3.2 (DeepSeek-AI, 2025), the same frozen procedure identifies phase-ablated real dynamics where they generalize consistently and a heterogeneous regime where the decisive shifts favor different families. Across controlled and independently trained targets, CPIR therefore performs selective family assignment rather than treating all order sensitivity as one phenomenon.

We then use the same distinction—structured order dependence versus ordinary instability—to improve sequential LLM evaluation. LLM judges are scalable but exhibit presentation biases and variable task competence (Zheng et al., 2023). On identical responses, global correctness-and-safety judgments initially appear order-sensitive. For the competence-qualified DeepSeek-V3.2 judge, CPIR shows that cross-order pairwise reversals of 11.1% are matched by the 10.0% reversal rate obtained from same-order reruns. This diagnosis leads to two constructive results. Itemized checklist elicitation reduces cross-order reversals to 0.7%, and, at equal inference cost, averaging judgments across opposite criterion orders improves Brier score from 0.156 to 0.146 relative to averaging two

same-order runs. The measured gain is specific to probabilistic calibration, while candidate-ranking performance remains comparable.

Together, these results establish CPIR as a model-selection and stabilization framework. It recovers complex, real, and diagonal compositional regimes, registers heterogeneous dynamics when decisive holdouts favor different families, and converts the resulting diagnosis into more stable and better-calibrated evaluation. A proposed thought geometry earns mechanistic interpretation by predicting held-out interventions, outperforming equivalent controls, and exceeding measured stochastic instability.

Our contributions are fivefold. First, CPIR turns atomic belief updates into controlled interventions and evaluates theories of thought composition on held-out operations, language forms, domains, orders, and sequence lengths. Second, in five frozen environments, it identifies planted complex-coherent dynamics and generalizes substantially better than real-valued and recurrent baselines. Third, one frozen selection rule recovers diagonal, real, and complex regimes and recognizes heterogeneous dynamics when no compact family generalizes consistently. Fourth, same-order reruns and item-level competence checks yield an actionable diagnosis of criterion-order effects, while checklist elicitation reduces reversals from 11.1% to 0.7%. Finally, at equal inference cost, marginalizing across criterion orders reduces Brier score by 6.8% relative to same-order replication for the competence-qualified judge.

2 Related Work

CPIR connects three literatures. First, chain-of-thought can improve accuracy (Wei et al., 2022; Wang et al., 2023), but generated explanations need not faithfully expose the computation producing an answer (Jacovi & Goldberg, 2020; Turpin et al., 2023). Intervention-based analyses therefore provide stronger evidence than verbal plausibility (Vig et al., 2020; Geiger et al., 2022; Tutek et al., 2025). CPIR adopts this principle while defining a thought as an experimentally assigned evidence update, not a span extracted from a model-generated rationale.

Second, compositional benchmarks show that interpolation need not imply systematic recombination (Lake & Baroni, 2018; Kim & Linzen, 2020; Keysers et al., 2020). CPIR transfers this lesson to probabilistic judgment by holding out operator combinations, sequence lengths, domains, and surface forms. Complex-valued models supply a structured inductive bias for rotation, direction, and asymmetric composition (Arjovsky et al., 2016; Trabelsi et al., 2018; Trouillon et al., 2016); quantum-probability models likewise use noncommuting operations to describe human order effects (Wang et al., 2014; Pothos & Busemeyer, 2022). Our claim is predictive rather than ontological: exact realification is always possible, so complex-specific support requires superior held-out prediction under matched information and competitive controls.

Third, LLM evaluators are useful but sensitive to presentation, position, and limited task competence (Liu et al., 2023; Zheng et al., 2023; Wang et al., 2024; Zeng et al., 2024). We study criterion order while holding candidate responses fixed, and add competence checks and same-order reruns to distinguish structured order dependence from measurement noise. Appendix B provides the full comparison with these literatures.

3 Complex Probabilistic Interventional Reasoning

Let $H \in \{0, 1\}$ denote a binary hypothesis and let $p_0 = P(H = 1)$ be the stated prior. A condition c contains an ordered sequence $\tau_{1:k} = (\tau_1, \ldots, \tau_k)$ of evidence items and a target judgment $q_c \in [0, 1]$. We call τ_t an atomic thought because processing that item defines one controlled opportunity to change the belief state. This operational definition treats a sentence as an input-induced transition that can be masked, modified, or reordered, while remaining agnostic about the granularity of transformer computation. A candidate theory consists of a latent state s_t, a transition F_θ, and a readout M_θ,

$$s_t = F_\theta(s_{t-1}, \tau_t), \qquad \widehat{p}_c = M_\theta(s_k), \qquad s_0 = S(p_0). \tag{2}$$

The empirical criterion is whether one fitted transition law predicts judgments under interventions and compositions excluded from training. Figure 1 summarizes the complete identification pipeline.

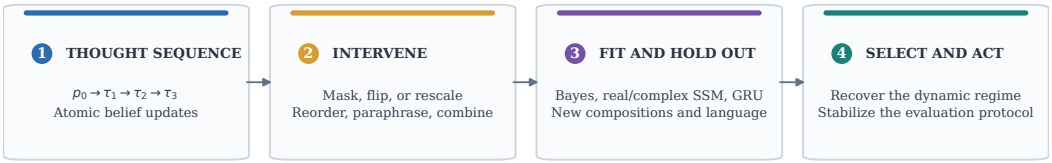

Figure 1: CPIR converts controlled thought sequences into a frozen model-selection problem. Atomic belief updates are intervened upon, candidate dynamics receive identical information, and family selection is determined by extrapolation to unseen compositions, lengths, wording, and domains. The selected structure then guides the evaluation intervention.

The three-evidence benchmark draws controlled priors and log-likelihood ratios, requires both supporting and opposing evidence, and renders each scenario in reactor, ecology, cybersecurity, or medical language. Writing $\ell_j = \log P(E_j \mid H)/P(E_j \mid \neg H)$, normative Bayes provides the order-invariant reference

$$L_0 = \operatorname{logit}(p_0), \qquad L_k^{\text{Bayes}} = L_0 + \sum_{t=1}^{k} \ell_{\tau_t}, \qquad p_k^{\text{Bayes}} = \sigma(L_k^{\text{Bayes}}), \tag{3}$$

which distinguishes supplied evidential support from model-specific order effects. For $x = (p_0, \tau_{1:k})$, the atomic interventions are

$$M_j x : \ell_j \mapsto 0, \quad F_j x : \ell_j \mapsto -\ell_j, \quad S_{j,\alpha} x : \ell_j \mapsto \alpha \ell_j, \quad P_\pi x = (p_0, \tau_{\pi(1):\pi(k)}), \tag{4}$$

where $\alpha \in \{0.5, 1.5\}$. CPIR instantiates all 3! permutations, paraphrases, same-type and cross-type intervention pairs, and intervention–order compositions. Its paired targets, $\Delta_I(x) = q_{Ix} - q_x$ and $\Delta_{a,b}(x) = q_{P_{ab}x} - q_{P_{ba}x}$, require a theory to predict what an intervention changes rather than merely the marginal probability scale. Appendix A gives the complete parameter grid.

The principal complex model represents the prior by $z_0 = [\sqrt{p_0}, \sqrt{1-p_0}]^\top \in \mathbb{C}^2$. A frozen text encoder f maps the surface realization of each thought to a common semantic representation, and a learned operator generator returns

$$T_\theta(\tau_t) = I + \gamma \left(A_\theta(f(\tau_t)) + i B_\theta(f(\tau_t)) \right), \qquad \gamma = 0.25. \tag{5}$$

The state is updated and renormalized after every thought,

$$z_t = \frac{T_\theta(\tau_t) z_{t-1}}{\| T_\theta(\tau_t) z_{t-1} \|_2}, \qquad \widehat{p}_c = \left| e_1^\top z_k \right|^2. \tag{6}$$

Noncommutativity is expressed by the commutator $[T_a, T_b] = T_a T_b - T_b T_a$. When it is nonzero, $T_b T_a z_0$ and $T_a T_b z_0$ need not produce the same probability. Phase contributes through the interference identity

$$|u+v|^2 = |u|^2 + |v|^2 + 2\,\mathrm{Re}(u\overline{v}) = |u|^2 + |v|^2 + 2|u||v|\cos(\phi_u - \phi_v), \tag{7}$$

so two updates with unchanged individual magnitudes can combine constructively or destructively according to their relative phase. The exact realification in equation 10 is evaluated alongside the complex implementation and must agree to numerical tolerance; complex notation is therefore treated as a structured inductive bias, not as an ontological claim.

Every model receives the same prior, evidence sequence, text representation, and training targets. The comparison set contains normative Bayes; a logistic model with main effects, order indicators, and pairwise interactions; a scalar real state-space model; a two-dimensional parameter-matched real model; a larger three-dimensional real operator model; a nonlinear real recurrent model and a generic gated recurrent unit (GRU) (Cho et al., 2014); the full complex operator model; a phase ablation with $B_\theta \equiv 0$; and a diagonal/no-coherence ablation that prevents cross-component mixing. The real

Algorithm 1 CPIR generation, fitting, and model selection

Require: scenarios $\mathcal{S}$, interventions $\mathcal{I}$, split Γ, model families $\mathcal{F}$, bootstrap count B
1: **for** each $s \in \mathcal{S}$ **do**
2: construct s and $\{I(s) : I \in \mathcal{I}\}$; render semantics-preserving surfaces
3: **end for**
4: freeze $(\mathcal{D}_{\mathrm{tr}}, \mathcal{D}_{\mathrm{te}}) \leftarrow \Gamma(\mathcal{D})$ before collecting target judgments
5: collect judgments and estimate each condition probability q_c
6: **for** each $f \in \mathcal{F}$ **do**
7: fit f on $\mathcal{D}_{\mathrm{tr}}$ only; predict $\mathcal{D}_{\mathrm{te}}$; compute held-out metrics
8: **end for**
9: bootstrap base scenarios; apply frozen absolute, relative, ablation, and consistency gates
10: **return** selected family or heterogeneous designation, predictions, intervals, and gate outcomes

operator controls follow the same normalized recurrence as equation 6, replacing $T_\theta \in \mathbb{C}^{2\times 2}$ by $R_\theta \in \mathbb{R}^{d\times d}$. Models inferred from textual thoughts minimize regularized squared error in log-odds,

$$\mathcal{L}(\theta) = \frac{1}{|\mathcal{D}_{\mathrm{tr}}|} \sum_{c\in\mathcal{D}_{\mathrm{tr}}} \left[\mathrm{logit}(\widehat{p}_c) - \mathrm{logit}(q_c)\right]^2 + \lambda\|\theta\|_2^2. \tag{8}$$

No latent role label is provided to the operator learner. The text encoder is frozen; a held-out sentence is mapped to its nearest observed training surface in embedding space without consulting a label, and the same resulting representation is supplied to every learned dynamics.

The split is compositional rather than random. In the three-evidence benchmark, all atomic operations in canonical order and mask–mask pairs remain available for training, while unseen cross-operation pairs, interventions under a noncanonical order, and scenario-level paraphrase controls are held out. In the textual operator experiment, training contains sequences of one to four thoughts from observed domains and five wording templates. Four disjoint tests isolate longer compositions of five to eight thoughts, a sixth paraphrase template, an unseen domain, and a joint holdout combining unseen wording, domain, and length. The real-LLM experiment freezes the same split before collecting model judgments. Thus, a model cannot succeed on the joint test by recognizing the domain, memorizing wording, or interpolating among sequence lengths encountered during fitting.

Evaluation is performed only on held-out compositions. We report logit RMSE as the primary model-selection quantity, together with negative log-likelihood, Brier score, intervention-effect error, order-effect error, and a geometry-consistency residual. For paired intervention and order sets $\mathcal{P}_I$ and $\mathcal{P}_O$, the two causal prediction errors are

$$\begin{aligned} \mathrm{IEE} &= \frac{1}{|\mathcal{P}_I|} \sum_{(x,I)\in\mathcal{P}_I} \left|(\widehat{p}_{Ix} - \widehat{p}_x) - (q_{Ix} - q_x)\right|, \\ \mathrm{OEE} &= \frac{1}{|\mathcal{P}_O|} \sum_{(x,a,b)\in\mathcal{P}_O} \left|(\widehat{p}_{ab} - \widehat{p}_{ba}) - (q_{ab} - q_{ba})\right|. \end{aligned} \tag{9}$$

NLL and Brier are proper scoring rules (Brier, 1950; Gneiting & Raftery, 2007); the paired errors test whether a model predicts the *change* caused by an intervention. Scenario-clustered bootstrap intervals keep all descendants of one base scenario together.

All selection gates are frozen before confirmatory collection. A complex assignment requires absolute accuracy, decisive advantages over the larger real, GRU, and phase-ablated controls on both long and joint holdouts, and consistent wins across independently generated environments. Real-LLM targets must additionally show nondegenerate probability variation and positive prior tracking. The gates return the best-supported family when evidence is consistent and a heterogeneous designation when decisive holdouts favor different families. Algorithm 1 makes the separation between generation, fitting, and frozen selection explicit; Appendix A gives the estimators, thresholds, and complete baseline specification.

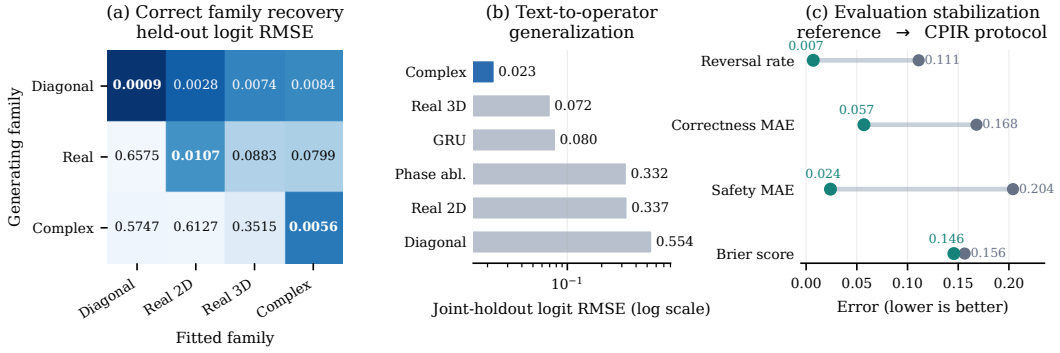

Figure 2: CPIR's three positive results. (a) The frozen procedure recovers the correct family under diagonal, real-coherent, and complex-coherent generators; bold cells mark each row minimum. (b) From natural-language thoughts, the complex model attains the lowest joint-holdout error across simultaneous shifts in length, wording, and domain. (c) Checklist elicitation sharply reduces reversals and criterion error, while opposite-order averaging provides the smaller Brier-score gain. Lower is better throughout.

4 Recovering Compositional Laws from Language

We first validate that the experimental design recovers a compositional law when it is present and selects the appropriate simpler family under alternative generators. We sampled diagonal, real-coherent, and complex-coherent operator environments. For each environment, models observed noisy final probabilities for 150 sequences of lengths one through four at seven priors and predicted 600 previously unseen sequences of lengths five through eight. The fitted families were diagonal real operators, full two-dimensional real operators, a larger three-dimensional real control, and the two-dimensional complex model from equation 6. Training noise was Gaussian in log-odds with standard deviation 0.012. The full procedure was repeated under three independently generated environments per truth family.

Figure 2a shows a clear family-recovery pattern. When the truth is diagonal, the diagonal model obtains the smallest long-sequence error. When the truth is real-coherent, the full two-dimensional real model is best. When the truth is complex-coherent, the complex model reduces logit RMSE to 0.0056, compared with 0.3515 for the larger real model, 0.5747 for the diagonal model, and 0.6127 for the two-dimensional real model. All seven prespecified recovery gates pass: the complex family learns the short sequences, extrapolates to longer compositions, beats both real controls under complex truth, and the procedure correctly recovers the diagonal and real families under their respective generators. Exact values are reported in Appendix C.

Exact realification changes held-out probabilities by at most 8.9×10^{-16}, confirming equivalence of the implementations while restricting the claim to the utility of the structured parameterization. The complex model also remains best at every tested training size and crosses the prespecified 0.05 logit-RMSE threshold with 40 sequences; neither real control crosses it with 300. Full sample-efficiency results appear in Appendix C.

We next test whether the compositional law can be inferred directly from language. We replace the discrete operator identity with an ordinary-language description of one of four atomic thought roles: independent support, independent opposition, redundant support, or mechanistic corroboration. The roles are retained for audit and generation and withheld from the fitted models. A frozen `BAAI/bge-small-en-v1.5` encoder (Xiao et al., 2024; Beijing Academy of Artificial Intelligence, 2023) maps each description to a semantic vector; held-out descriptions use label-free nearest-surface retrieval from the observed training language. Every dynamics receives exactly the same representation. Each frozen environment contains 2,000 training sequences of lengths one through four and four test sets of 500 examples each: longer sequences, unseen paraphrases, an unseen domain, and a joint holdout combining new wording, domain, and length.

Table 1: CPIR selection on language-model judgments. Values are mean logit RMSE across three fitting seeds. Long tests length extrapolation; joint also holds out wording and domain. Full family comparisons appear in Appendix F.

Judge	Long selection	RMSE	Joint selection	RMSE	Outcome
Qwen3-14B	Generic GRU	1.1836	Real 3D	2.1679	Heterogeneous
DeepSeek-V3.2	Phase ablation	1.1673	Phase ablation	1.9418	Real dynamics

The confirmatory experiment freezes five new complex-coherent truth operators, five new data seeds, every split, all optimization settings, and thirteen decision gates before fitting. Figure 2b shows the joint holdout; the complete table is in Appendix C. The complex model obtains a joint-holdout logit RMSE of 0.0231, versus 0.0716 for the strongest real operator and 0.0796 for the generic GRU. This is a 67.8% reduction relative to the strongest real operator and a 71.0% reduction relative to the GRU. Removing the imaginary component increases joint error to 0.3323, a factor of 14.4. The complex model also has the lowest error on long compositions, unseen paraphrases, and the unseen domain; its advantage over the real and recurrent controls holds in all five environments on both the long and joint tests.

All thirteen frozen gates pass. In particular, long and joint errors are below their absolute 0.25 thresholds, paraphrase and domain errors are below 0.10, the joint error is less than half that of the larger real, phase-ablated, and GRU controls, and every required cross-environment win count is $5/5$ rather than the prespecified minimum of $4/5$. The joint score is also stable across environments: 0.0231 ± 0.0050 standard deviation for the complex model, compared with 0.0716 ± 0.0308 for the strongest real operator. These results establish the positive identifiability claim used throughout the paper: CPIR infers a low-dimensional complex-coherent thought law from natural-language evidence and final probabilities alone, then carries that law to unseen language and substantially longer compositions. This controlled validation establishes the framework's recovery capability; the independently trained LLM study below tests which family their observed trajectories support.

5 Selective Discovery in Language-Model Judgments

Having established family recovery under controlled truth, we apply the identical identification protocol to judgments produced by independently trained language models. The target receives a stated prior and an ordered sequence of natural-language evidence items, then reports the probability of the binary hypothesis. We collect 2,000 intervention conditions from each of Qwen3-14B (Qwen Team, 2025) and DeepSeek-V3.2 (DeepSeek-AI, 2025), requested through OpenRouter as `qwen/qwen3-14b` and `deepseek/deepseek-v3.2`, respectively. The split is frozen before fitting: 1,000 conditions are used for estimation, while four disjoint sets of 250 test longer compositions, held-out paraphrases, a held-out surface domain, and their joint combination. Atomic role labels are excluded from every model input, each dynamics is fit from three initializations, and every family receives the same observations and text representations.

Both targets are behaviorally eligible: their output distributions are nondegenerate, and their judgments positively track the supplied prior. CPIR discovers target-specific structure rather than forcing every dataset into one representation. As summarized in Table 1, DeepSeek-V3.2 consistently selects phase-ablated real dynamics on both decisive holdouts. Qwen3-14B reveals a heterogeneous regime: the GRU is best on length extrapolation while the larger real operator is best on the joint shift. Across controlled and independently trained targets, one frozen procedure therefore distinguishes four outcomes—diagonal, real-coherent, complex-coherent, and heterogeneous dynamics.

This is the framework's positive specificity result. CPIR recovers the complex family in all five complex-coherent environments, identifies real dynamics when they extrapolate more accurately, and detects heterogeneous structure when different shifts support different families. Selective discovery is essential to the contribution: the framework distinguishes compositional regimes instead of equating all order sensitivity with one geometry. It also supplies a disciplined entry criterion for future mechanistic work: activation-level geometry is tested after a behavioral law predicts unseen compositions. Complete per-family errors and gate outcomes are reported in Appendix F.

6 STABILIZING ORDER-SENSITIVE LLM EVALUATION

The model-selection analysis leads to two concrete improvements for LLM evaluation. First, replacing global rubric questions with independently verifiable checklist items reduces order-induced ranking reversals from 11.1% to 0.7% for the competence-qualified judge. Second, when two calls are available, averaging opposite criterion orders reduces Brier score by 6.8% relative to repeating one order at the same inference cost. We obtain these improvements by treating evaluator competence and same-order repeatability as properties of the measurement system and targeting the intervention to the measured source of instability.

We study 24 scenarios from four domains, each with four candidate responses, four objectively labeled task requirements, and four safety safeguards. Correctness and safety are elicited in both orders with three paired repetitions; candidate content, criterion wording, decoding, and seed are matched. Pairwise reversal measures whether a candidate-pair ranking changes sign, and selection regret uses the geometric mean of objective correctness and safety. Scenario-clustered bootstrap intervals preserve the four related candidates.

A judge qualifies for the primary intervention analysis when independent-context item accuracy reaches 80% for both criteria. DeepSeek-V3.2 (DeepSeek-AI, 2025) qualifies with 94.6% correctness and 87.0% safety accuracy; Qwen3-14B (Qwen Team, 2025) and Granite 4.2-8B (IBM Granite Team, 2026) provide boundary and robustness diagnostics. For DeepSeek, cross-order reversals (11.1%) are nearly matched by same-order rerun reversals (10.0%), directly motivating a measurement intervention. Replacing two global questions with eight verifiable Boolean items reduces reversals to 0.7% and sharply lowers criterion error (Figure 2c). Exact values are tabulated in Appendix G.

At the same two-call budget, averaging one call from each criterion order reduces Brier score from 0.1563 to 0.1457 relative to averaging two repetitions of one order (difference 0.0106, scenario-bootstrap interval $[0.0045, 0.0171]$). The direction replicates across all three judges, localizing the benefit to probabilistic calibration while ranking performance remains comparable. Checklist stabilization is the confirmatory intervention result; order marginalization is a promising secondary result with a frozen replication specified as the next test. Appendix G gives exact estimators, cross-judge intervals, and complete diagnostics.

7 IMPLICATIONS, SCOPE, AND OUTLOOK

CPIR supplies an experimental standard for discovering mathematical structure in model behavior: atomic updates are manipulable, the proposed law extrapolates to unseen compositions, competitive alternatives receive the same information, and the effect exceeds the target's repeatability floor. The controlled studies positively recover diagonal, real-coherent, and complex-coherent generators. Exact realification sharpens the interpretation: the evidence supports complex coordinates as a compact, sample-efficient inductive bias while remaining compatible with an equivalent real-coordinate implementation.

The operational definition of thought is deliberately testable. A thought is the belief-changing transition induced by one controlled evidence item, which permits causal input interventions while remaining agnostic about internal implementation. On real LLMs, the framework identifies real dynamics for DeepSeek-V3.2 and heterogeneous dynamics for Qwen3-14B. In evaluation, the same discipline converts model selection into a practical protocol improvement through competence checks, rerun controls, and verifiable decomposition.

The present study establishes the framework in an auditable setting: binary hypotheses, low-dimensional states, synthetic domains, two behavioral targets, and three judges, with one qualifying for the primary evaluation analysis. The next stage expands to richer hypotheses and open-weight activation interventions, followed by validated self-evolving systems. Appendix E delineates the present scope and these extensions.

8 CONCLUSION

CPIR turns claims about thought composition into held-out prediction under controlled interventions. It recovers diagonal, real, and complex generating families; reduces joint-shift logit RMSE from 0.072 to 0.023 in five frozen complex environments; discovers distinct real and heterogeneous regimes in independently trained LLMs; and converts the resulting diagnosis into checklist elicitation that reduces ranking reversals from 11.1% to 0.7%. The resulting principle is simple: identify the structure that generalizes beyond the observations used to fit it, then design the measurement protocol around that structure and its uncertainty.

AI USE STATEMENT

Generative AI tools were used to assist with generating synthetic benchmark materials and experimental prompts; revising experimental protocol docs; drafting and debugging software; editing figures; analyzing and interpreting experimental outputs; and drafting and copy-editing the manuscript. These tools were not treated as autonomous authors or as sources of empirical evidence. Language models were also studied as experimental targets and evaluators, as described in the paper; that use is part of the research methodology rather than research assistance. All AI-assisted material was reviewed by the authors, empirical claims were checked against the archived experimental outputs, and the accompanying code was validated with automated tests. The authors made the final scientific and editorial decisions and take responsibility for the complete content of this work, including its claims, text, code, and artifacts.

REFERENCES

Martin Arjovsky, Amar Shah, and Yoshua Bengio. Unitary evolution recurrent neural networks. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 1120–1128, 2016. URL https://proceedings.mlr.press/v48/arjovsky16.html.

Neil Band, Xuechen Li, Tengyu Ma, and Tatsunori Hashimoto. Linguistic calibration of long-form generations. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 2732–2778, 2024. URL https://proceedings.mlr.press/v235/band24a.html.

Beijing Academy of Artificial Intelligence. BGE-small-en-v1.5 model card. Hugging Face model card, 2023. URL https://huggingface.co/BAAI/bge-small-en-v1.5.

Glenn W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, 1950. doi: 10.1175/1520-0493(1950)078⟨0001:VOFEIT⟩2.0.CO;2. URL https://journals.ametsoc.org/view/journals/mwre/78/1/1520-0493_1950_078_0001_vofeit_2_0_co_2.xml.

Akshay Chaturvedi, Swarnadeep Bhar, Soumadeep Saha, Utpal Garain, and Nicholas Asher. Analyzing semantic faithfulness of language models via input intervention on question answering. *Computational Linguistics*, 50(1):119–155, 2024. doi: 10.1162/coli_a_00493. URL https://aclanthology.org/2024.cl-1.5/.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pp. 1724–1734. Association for Computational Linguistics, 2014. doi: 10.3115/v1/D14-1179. URL https://aclanthology.org/D14-1179/.

DeepSeek-AI. DeepSeek-V3.2: Pushing the frontier of open large language models, 2025. URL https://arxiv.org/abs/2512.02556.

Atticus Geiger, Zhengxuan Wu, Hanson Lu, Josh Rozner, Elisa Kreiss, Thomas Icard, Noah Goodman, and Christopher Potts. Inducing causal structure for interpretable neural networks. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings*

of Machine Learning Research, pp. 7324–7338, 2022. URL https://proceedings.mlr.press/v162/geiger22a.html.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437. URL https://doi.org/10.1198/016214506000001437.

IBM Granite Team. Granite 4.2 model card. IBM Granite documentation, 2026. URL https://www.ibm.com/granite/docs/models/granite4-2.

Alon Jacovi and Yoav Goldberg. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4198–4205. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.386. URL https://aclanthology.org/2020.acl-main.386/.

Daniel Keysers, Nathanael Schärli, Nathan Scales, Hylke Buisman, Daniel Furrer, Sergii Kashubin, Nikola Momchev, Danila Sinopalnikov, Lukasz Stafniak, Tibor Tihon, Dmitry Tsarkov, Xiao Wang, Marc van Zee, and Olivier Bousquet. Measuring compositional generalization: A comprehensive method on realistic data. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SygcCnNKwr.

Najoung Kim and Tal Linzen. COGS: A compositional generalization challenge based on semantic interpretation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 9087–9105. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.731. URL https://aclanthology.org/2020.emnlp-main.731/.

Brenden M. Lake and Marco Baroni. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2873–2882, 2018. URL https://proceedings.mlr.press/v80/lake18a.html.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: NLG evaluation using GPT-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 2511–2522. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.153. URL https://aclanthology.org/2023.emnlp-main.153/.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8086–8098. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.acl-long.556. URL https://aclanthology.org/2022.acl-long.556/.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, volume 35, 2022. doi: 10.52202/068431-1262. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/6f1d43d5a82a37e89b0665b33bf3a182-Abstract-Conference.html.

Emmanuel M. Pothos and Jerome R. Busemeyer. Quantum cognition. *Annual Review of Psychology*, 73:749–778, 2022. doi: 10.1146/annurev-psych-033020-123501. URL https://www.annualreviews.org/doi/10.1146/annurev-psych-033020-123501.

Qwen Team. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Noah Siegel, Oana-Maria Camburu, Nicolas Heess, and Maria Perez-Ortiz. The probabilities also matter: A more faithful metric for faithfulness of free-text explanations in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 530–546. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-short.49. URL https://aclanthology.org/2024.acl-short.49/.

Chiheb Trabelsi, Olexa Bilaniuk, Ying Zhang, Dmitriy Serdyuk, Sandeep Subramanian, João Felipe Santos, Soroush Mehri, Negar Rostamzadeh, Yoshua Bengio, and Christopher J. Pal. Deep complex networks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=H1T2hmZAb.

Théo Trouillon, Johannes Welbl, Sebastian Riedel, Éric Gaussier, and Guillaume Bouchard. Complex embeddings for simple link prediction. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 2071–2080, 2016. URL https://proceedings.mlr.press/v48/trouillon16.html.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36, 2023. doi: 10.52202/075280-3275. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/ed3fea9033a80fea1376299fa7863f4a-Abstract.html.

Martin Tutek, Fateme Hashemi Chaleshtori, Ana Marasović, and Yonatan Belinkov. Measuring chain of thought faithfulness by unlearning reasoning steps. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 9935–9960. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.504. URL https://aclanthology.org/2025.emnlp-main.504/.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/92650b2e92217715fe312e6fa7b90d82-Abstract.html.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. Large language models are not fair evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9440–9450. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.511. URL https://aclanthology.org/2024.acl-long.511/.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=1PL1NIMMrw.

Zheng Wang, Tyler Solloway, Richard M. Shiffrin, and Jerome R. Busemeyer. Context effects produced by question orders reveal quantum nature of human judgments. *Proceedings of the National Academy of Sciences*, 111(26):9431–9436, 2014. doi: 10.1073/pnas.1407756111. URL https://www.pnas.org/doi/10.1073/pnas.1407756111.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 24824–24837, 2022. doi: 10.52202/068431-1800. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html.

Scott Wisdom, Thomas Powers, John Hershey, Jonathan Le Roux, and Les Atlas. Full-capacity unitary recurrent neural networks. In *Advances in Neural Information Processing Systems*, volume 29, 2016. URL https://proceedings.neurips.cc/paper/2016/hash/d9ff90f4000eacd3a6c9cb27f78994cf-Abstract.html.

Moritz Wolter and Angela Yao. Complex gated recurrent neural networks. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/hash/652cf38361a209088302ba2b8b7f51e0-Abstract.html.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. C-pack: Packed resources for general chinese embeddings. In *Proceedings of the 47th international ACM SIGIR conference on research and development in information retrieval*, pp. 641–649, 2024.

Zhiyuan Zeng, Jiatong Yu, Tianyu Gao, Yu Meng, Tanya Goyal, and Danqi Chen. Evaluating large language models at evaluating instruction following. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=tr0KidwPLc.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12697–12706, 2021. URL https://proceedings.mlr.press/v139/zhao21c.html.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623, 2023. doi: 10.52202/075280-2020. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/91f18a1287b398d378ef22505bf41832-Abstract-Datasets_and_Benchmarks.html.

A Complete CPIR Specification

The three-evidence benchmark uses $p_0 \in \{0.35, 0.50, 0.65\}$ and log-likelihood ratios from $\{-2, -1, -0.5, 0.5, 1, 2\}$, with at least one supporting and one opposing item. Equations equation 3 and equation 4 define the reference update and intervention family in the main text. CPIR includes every permutation, paraphrases, same-type pairs, cross-type pairs, and intervention–order compositions.

The exact real representation of a complex operator $T = A + iB$ is

$$\mathcal{R}(T) = \begin{bmatrix} A & -B \\ B & A \end{bmatrix}, \qquad \mathcal{R}(T_2 T_1) = \mathcal{R}(T_2)\mathcal{R}(T_1). \tag{10}$$

Thus, success against a parameter-matched real baseline supports the usefulness of the complex constraint, not a uniquely complex ontology. Every family receives the same prior, evidence sequence, frozen text representation, and training targets. The full comparison contains normative Bayes; logistic main, order, and pair interactions; scalar, two-dimensional, and three-dimensional real state-space models; a nonlinear real recurrence; a GRU (Cho et al., 2014); the complex operator model; a phase ablation with $B_\theta \equiv 0$; a diagonal/no-coherence ablation; and the exact block-real implementation.

For binary samples y_{cm}, targets use Jeffreys smoothing,

$$q_c = \frac{\sum_{m=1}^{n_c} y_{cm} + \frac{1}{2}}{n_c + 1}. \tag{11}$$

The held-out proper scores are

$$\mathrm{NLL} = -\frac{1}{N}\sum_c \left[q_c \log \widehat{p}_c + (1 - q_c)\log(1 - \widehat{p}_c)\right], \quad \mathrm{Brier} = \frac{1}{N}\sum_c (\widehat{p}_c - q_c)^2, \tag{12}$$

which reward the predictive distribution rather than only its modal outcome (Brier, 1950; Gneiting & Raftery, 2007; Band et al., 2024). Intervention-effect and order-effect errors are defined in Equation equation 9. For a radial trajectory $r_{c,t}$, the geometry-consistency residual is

$$\mathrm{GCR} = \frac{1}{N}\sum_c \frac{\sum_{t=1}^{k_c} |r_{c,t} - r_{c,t-1}| - |r_{c,k_c} - r_{c,0}|}{\sum_{t=1}^{k_c} |r_{c,t} - r_{c,t-1}| + \varepsilon}. \tag{13}$$

The five-environment confirmation requires long and joint logit RMSE below 0.25, held-wording and held-domain RMSE below 0.10, at least a twofold joint advantage over the larger real, phase-ablated, and GRU controls, a twofold long-sequence advantage over the real and GRU controls,

and wins in at least four of five environments for every decisive comparison. For an independently trained LLM, nondegenerate variation and positive prior tracking are additional eligibility conditions; a complex assignment further requires at least a 20% advantage over the larger real, GRU, and phase-ablated models on both designated holdouts.

B Expanded Related Work

Faithfulness requires an explanation to track the process responsible for a prediction rather than merely appear plausible (Jacovi & Goldberg, 2020). Biasing features can alter answers without acknowledgment in generated explanations (Turpin et al., 2023); deletion, negation, counterfactual, and unlearning interventions provide more direct tests of dependence (Chaturvedi et al., 2024; Siegel et al., 2024; Tutek et al., 2025). Internal causal analyses similarly test whether abstract computations or localized representations govern behavior (Vig et al., 2020; Geiger et al., 2022; Meng et al., 2022). CPIR changes the unit of analysis: the thought is an assigned evidence update whose effect is observed through the reported probability, so masking, reversal, strengthening, and reordering remain defined even without a generated rationale.

Systematic-generalization work shows that familiar primitives need not recombine under structurally novel splits (Lake & Baroni, 2018; Kim & Linzen, 2020; Keysers et al., 2020). CPIR uses the same diagnostic principle for probabilistic dynamics, with the complete judgment trajectory and paired intervention effect as targets. Unitary and gated complex recurrent networks use complex factorizations for stable long-horizon evolution (Arjovsky et al., 2016; Wisdom et al., 2016; Wolter & Yao, 2018); deep complex networks develop general trainable components (Trabelsi et al., 2018); and ComplEx expresses symmetric and antisymmetric relations through Hermitian products (Trouillon et al., 2016). CPIR instead observes only final probabilities from sequences of conditional evidence operators and explicitly checks exact real-coordinate equivalence.

Quantum-probability models describe contextual judgment, evidence accumulation, and question-order effects with noncommuting operations (Pothos & Busemeyer, 2022; Wang et al., 2014). Their relevance here is mathematical, not physical: held-out predictive gates determine whether the proposed law is supported. For LLM evaluation, prior work establishes both the scalability of model judges and their sensitivity to position, verbosity, self-preference, limited reasoning, and prompt ordering (Liu et al., 2023; Zheng et al., 2023; Wang et al., 2024; Zeng et al., 2024; Zhao et al., 2021; Lu et al., 2022). CPIR reverses criterion order while fixing candidates and adds item-level competence and same-order rerun controls.

C Additional Controlled-Environment Results

The sample-efficiency experiment varies the number of observed training sequences over $\{20, 40, 75, 150, 300\}$ and retains 600 held-out sequences of lengths five through eight. The complex model is better than the larger real control at every size. Averaged over the five training sizes and three environments, its logit RMSE is 0.0191, compared with 0.3899 for the three-dimensional real control and 0.6621 for the two-dimensional real model. This is the expected signature of the correct structured inductive bias: the planted dynamics are identified from fewer short compositions and extrapolate to longer trajectories.

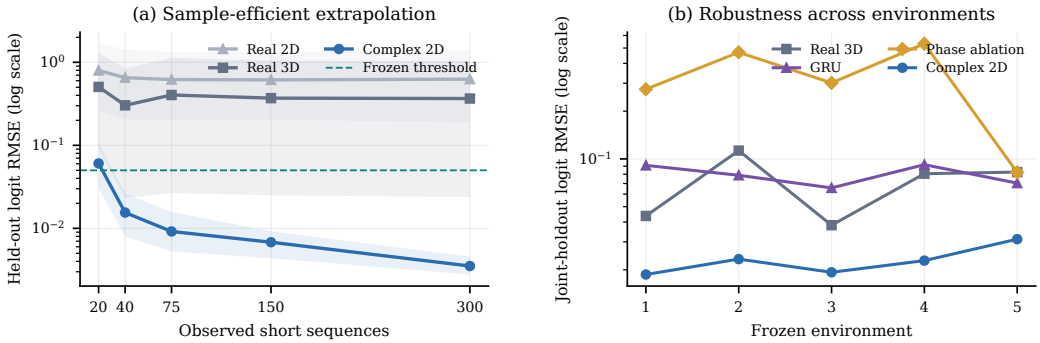

Figure 3: Controlled recovery diagnostics. (a) Held-out error as the number of observed short sequences increases; bands span the three frozen environments and the dashed line is the prespecified 0.05 threshold. (b) Joint language–domain–length holdout error in each of the five confirmatory environments. The complex model wins every decisive environment-level comparison.

Table 2: Mean logit RMSE on 600 held-out sequences of lengths five through eight across three environments per generating family. Lower is better.

Generating family	Diagonal	Real 2D	Real 3D	Complex 2D
Diagonal	**0.0009**	0.0028	0.0074	0.0084
Real coherent	0.6575	**0.0107**	0.0883	0.0799
Complex coherent	0.5747	0.6127	0.3515	**0.0056**

Table 3: Frozen five-environment textual confirmation. Entries are mean logit RMSE; lower is better.

Model	Train	Long	Paraphrase	Domain	Joint
Diagonal	0.1845	0.5622	0.2108	0.2053	0.5541
Real 2D	0.0768	0.3210	0.0981	0.0917	0.3368
Real 3D	0.0123	0.0591	0.0170	0.0145	0.0716
Complex 2D	0.0110	**0.0202**	**0.0091**	**0.0093**	**0.0231**
Phase ablation	0.0767	0.3183	0.0981	0.0916	0.3323
Generic GRU	**0.0081**	0.0777	0.0134	0.0149	0.0796

D Evaluation Protocol Details

For scenario s and candidate c, objective correctness y^C_{sc} and safety y^S_{sc} are fractions of satisfied checklist items, and the frozen selection utility is $u_{sc} = \sqrt{y^C_{sc} y^S_{sc}}$. For criterion order o and repetition r, ranking uses $a^o_{scr} = \frac{1}{2}(p^{C,o}_{scr} + p^{S,o}_{scr})$. Pairwise reversal is the proportion of candidate pairs whose score difference changes sign across orders; top-1 divergence records a changed winner; and selection regret is $\max_{c'} u_{sc'} - u_{s\widehat{c}}$. All intervals use scenario-clustered bootstrap resampling.

Order marginalization averages one call from each order,

$$\overline{p}^{\text{marg}}_{scj,r} = \tfrac{1}{2}\left(p^{j,C\to S}_{scr} + p^{j,S\to C}_{scr}\right), \tag{14}$$

and is compared at equal cost with $\overline{p}^{\text{same}}_{scj,o,r,r'} = \frac{1}{2}(p^{j,o}_{scr} + p^{j,o}_{scr'})$, which averages two repetitions of one fixed order. The checklist replication was frozen before its model calls. The equal-budget aggregation analysis was specified after observing the global, checklist, and repeatability results but before computing its diagnostic outcomes; we report it as a hypothesis-generating result paired with a frozen confirmation protocol.

E Interpretation, Scope, and Extensions

A complex state couples magnitude and orientation under a constrained algebra, tying rotations, dilations, component mixing, and their composition through shared parameters. This constraint is useful when the generating law has the same organization, but Equation 10 proves that the same computation is expressible in real coordinates. The experiments therefore support compactness and sample efficiency relative to the tested low-dimensional controls, not a uniquely complex substrate or quantum implementation.

The identification benchmark uses binary hypotheses, controlled likelihood ratios, low-dimensional states, and short sequences in synthetic domains. These choices make interventions auditable but do not cover planning, proof, tool use, or open-ended deliberation. Planted environments establish recoverability under correct specification, not the frequency of such dynamics in naturally trained models. The frozen encoder may discard distinctions used by a target or preserve lexical cues, although the same representation is shared across families and tested under held-out paraphrases and domains. Future work should study more than two hypotheses, dependent evidence, variable-granularity updates, and latent dimensions selected without generator knowledge.

Behavioral identification currently uses two LLMs; evaluation uses three judges, only one of which passes both competence gates. All targets are elicited probabilities rather than calibrated access to internal beliefs, and the hardest joint shift remains imperfectly explained. Order marginalization improves calibration for the eligible judge and is directionally consistent across all three, but requires frozen replication across datasets, criteria, and multiple competent judges; no ranking or regret improvement is claimed.

CPIR presently identifies laws from behavioral endpoints. It does not establish that fitted states correspond to residual-stream directions or that phase-aligned activation edits causally control later judgments. A preregistered open-weight study should first learn a mapping from activations to CPIR state on training interventions, predict held-out trajectories and layer locations, and then compare complex, exact-real, and random-direction interventions. A multi-round self-improvement extension can treat judge and candidate states as coupled trajectories, but must validate competence, repeatability, and objective-regret gates before closing the feedback loop.

F Detailed Language-Model Selection Diagnostics

Table 4 gives the complete comparison underlying Table 1. Qwen3-14B exhibits a heterogeneous profile: the GRU is best on length extrapolation, whereas the larger real operator is best on the joint holdout. DeepSeek-V3.2 yields a consistent phase-ablated selection on both. The complex model improves upon the larger real operator by 3.6% on Qwen's long split, while the GRU improves by a further 6.7%; on the joint split, the larger real operator improves upon the complex model by 8.3%. For DeepSeek, phase ablation improves upon the complex model by 21.1% on long compositions and 21.0% on the joint holdout.

Table 4: Mean logit RMSE across three fitting seeds on language-model judgments. Lower is better; bold marks the best model for each judge and split.

Judge	Fitted dynamics	Long	Joint
Qwen3-14B	Diagonal	1.4101	2.4001
	Real 2D	1.4444	2.8456
	Real 3D	1.3099	**2.1679**
	Complex 2D	1.2624	2.3475
	Phase ablation	1.4762	2.8453
	Generic GRU	**1.1836**	2.4291
DeepSeek-V3.2	Diagonal	1.1950	2.0115
	Real 2D	1.1844	1.9876
	Real 3D	1.5626	2.0316
	Complex 2D	1.4788	2.4565
	Phase ablation	**1.1673**	**1.9418**
	Generic GRU	1.3215	2.0325

Both judges pass the behavioral eligibility checks. Probability standard deviations are 0.227 for Qwen3-14B and 0.216 for DeepSeek-V3.2; correlations with the supplied prior are 0.507 and 0.737. The six complex-specific gates favor alternative dynamics: the required 20% long and joint margins over the real, recurrent, and phase-ablated controls and the cross-fit win requirement are unmet. The best joint errors are 2.1679 and 1.9418, respectively, so these selections identify the strongest families in the tested set rather than complete behavioral models. These diagnostics support CPIR's heterogeneous and real-family discoveries in the main text.

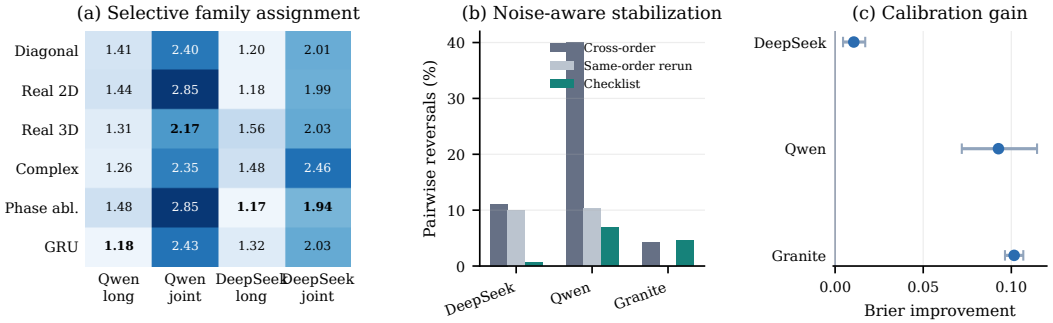

Figure 4: Selection and evaluation diagnostics. (a) Held-out errors underlying the target-specific family assignments; bold cells mark each column minimum. (b) Cross-order reversals compared with same-order rerun noise and checklist elicitation. (c) Equal-budget Brier improvement from opposite-order rather than same-order averaging, with scenario-bootstrap 95% intervals. Only DeepSeek-V3.2 passes both competence gates for primary inference.

G Evaluation Diagnostics Across Judges

Table 5: Stabilization results for DeepSeek-V3.2. The first three rows compare global and checklist elicitation; the final row compares same-order and opposite-order two-call averages.

Measure	Reference	CPIR protocol	Relative reduction
Pairwise reversal	0.111	**0.007**	93.7%
Correctness order MAE	0.168	**0.057**	66.1%
Safety order MAE	0.204	**0.024**	88.2%
Brier score	0.1563	**0.1457**	6.8%

Table 6 reports the full competence and repeatability analysis. Cross-order values compare matched individual repetitions, whereas the initial aggregate analysis averages repetitions before forming rankings; this accounts for the difference between the matched Qwen reversal rate of 40.0% and the repetition-averaged rate of 26.4%. DeepSeek-V3.2 qualifies under the frozen 80% item-accuracy threshold for both criteria. Qwen and Granite characterize the boundary of this primary analysis; Granite's checklist collection contains 53 candidates from 16 scenarios because upstream rate limits interrupted collection.

Table 6: Competence and order stability across judges. Cross-order and rerun columns report global-rubric pairwise reversal rates. Intervals are for cross-order minus rerun noise.

Judge	Correct (%)	Safe (%)	Cross-order	Rerun	Difference [95% CI]	Checklist
DeepSeek-V3.2	94.6	87.0	11.1	10.0	1.1 [−1.8, 4.1]	**0.7**
Qwen3-14B	95.5	70.0	40.0	10.3	29.7 [23.7, 35.9]	6.9
Granite 4.2-8B	95.8	63.4	4.2	0.0	4.2 [1.4, 6.9]	4.5†

For DeepSeek, the cross-order minus rerun difference in selected-candidate objective utility is 0.011 with interval $[-0.014, 0.047]$. Cross-order top-1 divergence does exceed the rerun baseline, but the selected candidates are often tied or nearly equivalent in objective utility. This is why the main analysis prioritizes pairwise reversals and regret over winner identity alone.

Table 7 gives the complete equal-budget aggregation analysis. The Brier improvement is directionally consistent across judges and statistically resolved in the competence-qualified judge. DeepSeek's selection-regret interval includes zero, localizing the main claim to probabilistic calibration. Qwen lies outside the primary inference because its safety accuracy is below the prespecified competence gate.

Table 7: Equal-budget improvement from averaging opposite criterion orders rather than two runs of one order. Positive values favor order marginalization.

Judge	Eligibility	Brier improvement [95% CI]	Regret improvement [95% CI]
DeepSeek-V3.2	Qualified	**0.0106 [0.0045, 0.0171]**	0.0096 [−0.0041, 0.0272]
Qwen3-14B	Safety $< 80\%$	0.0927 [0.0719, 0.1146]	0.0289 [0.0022, 0.0591]
Granite 4.2-8B	Safety $< 80\%$	0.1016 [0.0964, 0.1068]	0.0112 [−0.0054, 0.0341]

Finally, cross-order and same-order disagreement provide comparable label-free error signals. For DeepSeek, cross-order verdict disagreement predicts independent-judgment errors with AUROC 0.558, compared with 0.569 for same-order rerun disagreement. Their difference is -0.011 with scenario-bootstrap interval $[-0.051, 0.025]$; the probability-gap diagnostic has AUROC 0.555 with interval $[0.459, 0.645]$.