Lost in Projection: The Geometric Orthogonality of LLM Assessment and Reasoning

Anonymous ACL submission

Abstract

Large language models (LLMs) frequently exhibit a dissociation between their internal confidence representations and actual performance. Identifying the mechanistic origins of this gap is challenging, as representational analyses are often confounded by surface-level heuristics and fail to disentangle internal assessment from procedural execution. In this work, we characterize a dual-subspace structure within the residual stream, demonstrating that the pre-generative representational assessment of task difficulty is nearly orthogonal to subsequent generation dynamics. Using length-controlled datasets and statistical hypothesis testing against randomized baselines, we decode a pre-generative representation of task solvability across three model families (Llama, Qwen, Mistral). However, while causal steering within this high-dimensional *Assessment Subspace* successfully shifts the decoded confidence ($\Delta > 0.8$), it has no statistically significant effect on downstream reasoning accuracy even on borderline tasks where the model possesses the requisite capability. We show that this decoupling coincides with a dimensionality shift at the onset of token generation, where the activation trajectory transitions into a distinct, low-rank *Execution Subspace*. These results suggest that pre-generative confidence representations do not serve as causal control levers for procedural competence, indicating that future reliability frameworks should leverage these early assessment states for dynamic compute routing rather than behavioral steering.

1 Introduction

A challenge in the deployment of Large Language Models (LLMs) (Grattafiori et al., 2024; Jiang et al., 2023; Qwen et al., 2025) is their tendency to generate contextually coherent but factually incorrect responses, often while maintaining high represented confidence (Maharana et al., 2025; Xiong et al., 2023; Ren et al., 2023). This dissociation between represented confidence and task competence poses a reliability problem for applications requiring multi-step reasoning, such as scientific discovery and medical diagnostics. While prior research has documented this confidence-competence gap behaviorally (Singh et al., 2023) and noted the inconsistency of activation steering methods (Tan et al., 2025; Braun et al., 2025), the mechanistic origin of this phenomenon remains unresolved. This paper investigates why the pre-generative representation of a problem's solvability is functionally decoupled from the subsequent generation process, providing a characterization based on the geometry of representations.

A prevailing hypothesis in mechanistic interpretability suggests that identifying the latent representation of a model's expected success—often termed its "solvability representation"—could allow for the direct control of its performance (Marks and Tegmark, 2024a; Dunefsky and Cohan, 2025). However, testing this hypothesis requires careful experimental control to isolate the semantic signal of task assessment from high-dimensional noise. Recent audits of interpretability methods warn that linear probes can easily align with high-dimensional noise or surface-level heuristics, leading to spurious causal claims (Méloux et al., 2025; Ravichander et al., 2021). To isolate the semantic representation of task assessment from these familiarity confounds, we employ a controlled framework using length-balanced datasets of multi-step reasoning problems across three model families (Qwen-2.5 7B (Qwen et al., 2025), Llama 3.2 3B (Grattafiori et al., 2024), Mistral Small 24B (Jiang et al., 2023)). We further validate our probes using statistical hypothesis testing against randomized computational baselines, providing statistical evidence that they capture a learned representation of solvability.

Our investigation proceeds in three stages. First,

we decode a latent, pre-generative assessment of task solvability. Using linear probes and Centered Kernel Alignment (CKA) (Kornblith et al., 2019b; Zhou et al., 2024; Murphy et al., 2024), we demonstrate that this assessment is linearly decodable prior to the onset of generation across diverse domains, including mathematics, coding, and logic. Second, we subject this state to a causal test. We intervene using targeted steering vectors to shift the hidden representation from *unsolvable* to *solvable* (inducing a probe-predicted probability mass shift of $\Delta > 0.8$) (Li et al., 2023a; Rimsky et al., 2023; Lee et al., 2024; Turner et al., 2023). We find that despite this large representational shift, downstream reasoning accuracy remains statistically unchanged, a null effect that persists even under stochastic decoding regimes. Finally, to explain this functional decoupling, we analyze the geometry of the activation space. Using principal component analysis and cross-subspace projection, we identify a geometric transition (Park et al., 2023; Marks and Tegmark, 2023; Balestriero et al., 2023; Konen et al., 2024) between the high-dimensional manifold of the **Assessment Subspace** and the low-dimensional manifold of the **Execution Subspace**. We show that these subspaces are nearly orthogonal and separated by an effective dimensionality shift at the onset of token generation, isolating the initial assessment from the subsequent dynamics of execution.

The characterization of this dual-subspace structure suggests a structural limit to representation engineering and a recalibration of strategies for AI reliability. The implications of our findings are summarized as follows:

- **Dynamic Compute Routing:** While the pre-generative Assessment signal is causally inert for behavioral control, it remains predictive. This offers a practical pathway for efficient NLP inference: early signals from the Assessment Subspace can trigger dynamic compute allocation (e.g., routing to larger models) or early exiting when the model predicts a high likelihood of failure, saving computational resources on exhaustive generation.

- **Causal Non-Identifiability of Confidence:** Increasing a model's internal confidence representation is decoupled from reasoning reliability in our experiments. The bidirectional near-orthogonality between assessment and execution indicates that pre-generative evaluation states are not causal control levers for procedural tasks.

- **Focus on Procedural Dynamics:** Addressing reasoning failures suggests a shift from high-level interventions on abstract traits (like "confidence" or "honesty") to methods that directly target the low-rank dynamics of the execution subspace during active generation.

- **Mechanistic Auditing:** Conventional benchmarks focused solely on final answers provide an incomplete evaluation of model reliability. Future evaluations should independently audit the predictive fidelity of the Assessment Subspace and the procedural robustness of the Execution Subspace.

2 Setup

To ensure our findings reflect general principles of LLM architecture rather than idiosyncratic model behaviors, we evaluate a diverse panel of instruction-tuned models from distinct organizations: **Qwen 2.5 7B** (Qwen et al., 2025), **Llama 3.2 3B** (Grattafiori et al., 2024), and **Mistral Small 24B** (Jiang et al., 2023). This selection allows us to validate the observed decoupling across varying parameter scales and design philosophies.

We subject these models to a suite of reasoning tasks spanning three distinct cognitive domains: (i) **Numerical Reasoning** (GSM8K (Cobbe et al., 2021), Math-Hard (Hendrycks et al., 2021), OpenR1 Math (Zhao et al., 2025)), (ii) **Formal Logical Deduction** (Knights and Knaves (Xie et al., 2024)), and (iii) **Algorithmic Planning** (Open R1 Coding (Zhao et al., 2025), QWQ-Planning (Hook, 2025)). This broad task coverage ensures that the dissociation between assessment and execution is a fundamental property of the models' reasoning processes rather than an artifact of a specific domain. To maintain methodological control, we also include a suite of trivial, non-reasoning Control Tasks (e.g., verbatim copying and rote sequence retrieval) to establish baseline dimensionality metrics. All experiments were performed on a compute cluster of three NVIDIA A6000 GPUs; see Appendix C.2.1 for comprehensive dataset descriptions, prompt templates, and hardware specifications.

3 The Anatomy of Pre-Generative Assessment

To test whether hidden states encode a pre-generative assessment of task success, we probe the residual stream at the final prompt token. We isolate this representation from superficial heuristics using a curation protocol that strictly controls for prompt length and formatting confounds.

3.1 A Protocol for Confound-Resistant Data Curation

To isolate the semantic representation of solvability from superficial heuristics (Lysnæs-Larsen et al., 2025; Kumar et al., 2022), we implement a multi-stage filtering protocol on 3,000 mathematical reasoning problems, sequentially neutralizing structural formatting, topic imbalance, and prompt length. This process yields a curated dataset of 846 examples with statistically indistinguishable token count distributions between the "solved" and "unsolved" classes ($p > 0.4$; see Appendix C.3.1). To prevent representational comparison bias, all diagnostic baselines (Logistic Regression, SVC, XGBoost, and MLP) are optimized with equal computational effort using a 5-fold grid-search cross-validation protocol over key hyperparameters (Table 9). This combined pipeline isolates a latent, semantic assessment state from superficial input artifacts while ensuring representational performance gaps are not confounded by suboptimal optimization.

3.2 Isolating and Characterizing the Latent Assessment State

Consistent with probing protocols established for autoregressive models (Tillman and Mossing, 2025; Kantamneni et al., 2025; Yan, 2025), we isolate the pre-generative assessment state by extracting hidden representations at the final token of the input prompt. This locus captures the model's fully contextualized representation of the problem immediately prior to the onset of the reasoning chain. To characterize the geometry of this representation, we investigate whether the solvability signal is encoded via a linear subspace or a complex non-linear manifold.

We address this by comparing the decoding performance of a linear probe against a hierarchy of non-linear classifiers. Specifically, we train a **Logistic Regression** probe to test for linear separability, and compare its performance to three non-linear baselines: a **Support Vector Classifier** with an RBF kernel, an **XGBoost** classifier, and a **2-Layer MLP** (hyperparameters detailed in Appendix C.4.1). This comparative framework serves as a controlled test of representational complexity: if the representation of task success relies on complex feature interactions, the non-linear models should significantly outperform the linear baseline.

The decoding results, presented in Figure 1, reveal two structural properties of this representation. First, solvability is linearly decodable, with accuracy peaking between **70–75%** in the mid-to-late layers across all model families. Second, the signal is fundamentally linear. As shown in Figure 1, the capacity-rich non-linear probes yield no statistically significant performance improvement over the simple logistic regression model. This indicates that the information regarding task solvability is accessible via a linear readout (see Appendix E.1). To confirm that our linear probe is reading a pre-existing semantic structure rather than fitting to high-dimensional representation noise (Méloux et al., 2025), we evaluate our probe against randomized computational baselines (Table 5). The probe collapses to chance ($\sim$50%) under both random initialization and shuffled labels, verifying that the solvability feature is a trained, stable property of the representation space.

To validate that our curated dataset size ($N = 846$) provides sufficient statistical power and to rule out memorization of superficial artifacts, we conduct a sample complexity analysis across all model families (detailed in Appendix F, Figure 7). We observe a distinct, three-stage representational phase transition as training sample size scales. In the *noise regime* ($N < 200$), probe accuracy remains at chance, suggesting the absence of simple lexical shortcuts or prompt-length correlates that would allow rapid, few-shot convergence. In the *crystallization phase* ($200 \leq N \leq 350$), generalization accuracy undergoes a geometric phase transition, rising sharply as the signal-to-noise ratio of the semantic direction overcomes the isotropic noise of the ambient space. Finally, in the *stable regime* ($N > 350$), performance stabilizes and plateaus asymptotically between 70–75%. This stable plateau mathematically confirms that our chosen dataset size resides securely within the converged regime, ensuring that our derived steering vectors (d_{solv}) represent robust, asymptotic estimates of the underlying solvability manifold.

Formally, let $\mathcal{H} \subset \mathbb{R}^d$ denote the hidden state

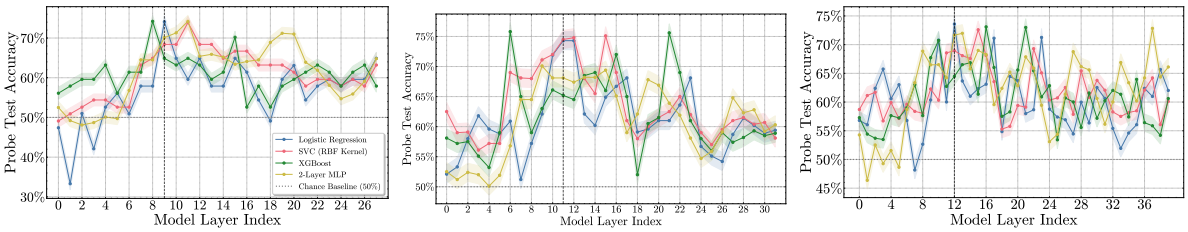

Figure 1: **Linear Decodability of the Assessment State.** Probe accuracy across network layers for three model families (**Left to Right: Qwen 2.5 7B, Llama 3.2 3B, Mistral Small 24B**). Despite the increased expressive capacity of non-linear probes (SVC, XGBoost, MLP), they fail to outperform the linear baseline (Logistic Regression). This structural equivalence indicates that the solvability representation is encoded as a linear direction within the residual stream, emerging robustly in the mid-to-late layers.

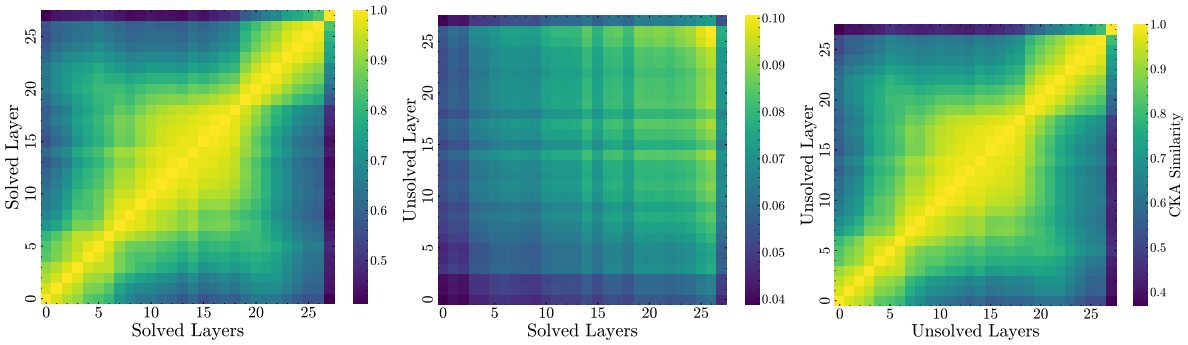

Figure 2: **Geometric Coherence and Dissimilarity of Assessment States**. Using Centered Kernel Alignment (CKA), we compare the representational geometry of assessment states. **(Left & Right)** The high self-similarity along the diagonals suggests that both "Solved" and "Unsolved" states are internally coherent, geometrically stable representations across layers. **(Center)** In stark contrast, the cross-comparison reveals a profound geometric dissimilarity, with near-zero CKA scores between the two states. This provides strong evidence that the model represents assessment not as a single continuum, but as two distinct and fundamentally separate geometric objects.

space at a given layer. For a problem i, let $h_i \in \mathcal{H}$ be its activation vector and $y_i \in \{0, 1\}$ its ground-truth solvability label. The equivalence in performance between linear and non-linear probes implies the existence of a specific direction $d_{\text{solv}} \in \mathbb{R}^d$ and bias b such that the probability of success is approximated by

$$P(y_i = 1 \mid h_i) \approx \sigma(h_i \cdot d_{\text{solv}} + b). \qquad (1)$$

While the signal is linearly separable, the imperfect accuracy ($\sim 75\%$) suggests that the projected distributions $P(h_i \cdot d_{\text{solv}} \mid y_i = 1)$ and $P(h_i \cdot d_{\text{solv}} \mid y_i = 0)$ maintain non-trivial overlap. This characterizes the assessment state as a linear direction embedded within a high-dimensional, noisy manifold, rather than a perfectly disentangled feature.

3.3 Geometry Visualization and Structural Similarity

To validate that the statistical separability observed via linear probing corresponds to a robust topological feature of the activation manifold, we visualize the high-dimensional hidden states using non-linear dimensionality reduction. We employ t-distributed Stochastic Neighbor Embedding (t-SNE) to examine local structural preservation and Uniform Manifold Approximation and Projection (UMAP) to assess global geometry.

As illustrated in Figure 3, which visualizes $N = 800$ data points, both projection methods reveal a clear geometric separation. The activation vectors resolve into disjoint clusters rather than a diffuse cloud. We note that the visual density of the clusters—appearing as compact groups despite the large sample size—results from significant overplotting, reflecting the high topological cohesion of the respective manifolds. This suggests that the solvability assessment is encoded in a distinct region of the activation space, consistent with the linear separability results.

To quantify the representational consistency and mutual distinctness of these states across layers, we calculate Centered Kernel Alignment (CKA) (Ko-

Table 1: **Probe Accuracy by Token Position.** We trained our suite of probes on activations extracted from different positions within the input prompt to identify the locus of the pre-generative assessment signal. Accuracy is reported on the held-out test set of our length-controlled math dataset. For all probe architectures, performance consistently and decisively peaks at the **Last Input Token** ('t_question_end'), indicating that the model's solvability representation crystallizes at the moment it has finished processing the full problem context.

Activation Locus	Logistic Reg.	SVC	XGBoost	MLP
10% of Prompt	54.2% $\pm$ 2.1	55.1% $\pm$ 2.3	53.8% $\pm$ 2.5	52.9% $\pm$ 2.8
50% of Prompt	65.1% $\pm$ 1.8	66.3% $\pm$ 1.9	65.9% $\pm$ 2.0	66.8% $\pm$ 2.1
Last Input Token	**74.4% $\pm$ 1.5**	**74.6% $\pm$ 1.6**	**75.1% $\pm$ 1.7**	**75.2% $\pm$ 1.8**
EOS Token	67.9% $\pm$ 1.6	68.5% $\pm$ 1.7	68.1% $\pm$ 1.8	68.8% $\pm$ 1.9

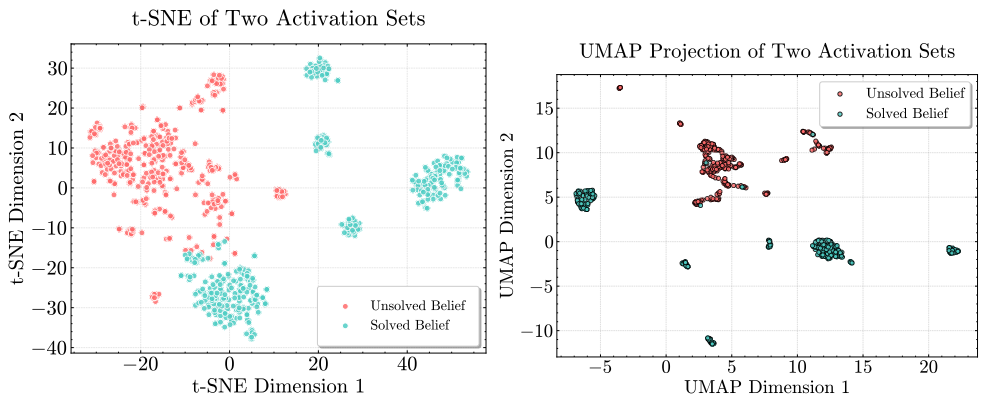

Figure 3: **Geometric Separation of Assessment States.** Visualization of high-dimensional assessment representations projected via t-SNE (**Left**) and UMAP (**Right**) for Qwen 2.5 7B (Layer 26, $N = 846$). Both manifold learning techniques demonstrate a distinct topological separation between "Unsolved" (coral) and "Solved" (cyan) classes. The emergence of well-defined clusters suggests that the statistical separability observed in Section 3.2 corresponds to a robust geometric structure within the activation space, rather than an artifact of high-dimensional sparsity.

rnblith et al., 2019a). The similarity heatmaps (Figure 2) demonstrate three representational properties. First, we observe intra-class coherence, where the self-similarity matrices for both "Solved" and "Unsolved" states exhibit stable, high CKA values along and near the diagonal across the mid-to-late layers, suggesting that these representations evolve consistently and maintain stability throughout the network. Second, the cross-comparison between "Solved" and "Unsolved" layers yields near-zero CKA scores, indicating that these states occupy geometrically distinct subspaces. Finally, this representational separation undergoes a layer-wise stabilization, emerging and crystallizing in the mid-to-late layers where the distinction between solvability classes remains preserved.

These analyses suggest the existence of a stable, geometrically distinct assessment manifold. However, geometric separability does not inherently imply functional utility. Having established the structural properties of this state, we next investigate its causal role in the reasoning process.

4 Causal Decoupling

4.1 Causal Interventions on the Latent Assessment Subspace

To investigate the causal role of the pre-generative assessment representation, we derive a steering vector, d_{solv}, normal to the decision boundary of our trained logistic regression probe. Trajectory analysis (Table 1) shows that linear decodability peaks at the final prompt token, consistent with prior protocols (Tillman and Mossing, 2025; Kantamneni et al., 2025; Yan, 2025), designating this token as our target intervention locus.

We apply an additive update to the residual stream at the last input token of layer L (Gurnee and Tegmark, 2024):

$$h'_{i,L} = h_{i,L} + \alpha \cdot d_{\text{solv}}, \tag{2}$$

where the scalar shift coefficient α is tuned via leave-one-out cross-validation (Hastie et al., 2009).

To test if the decodable representation functions as an active control variable for competence, we steered representations from *unsolvable* to *solvable* across four domains (Table 2). On Math-Hard, the intervention shifts the probe's predicted probability from 0.04 to 0.97 ($\Delta = +0.93$, representing the net change in predicted probability mass), yet downstream accuracy remains statistically unchanged ($0.0\% \pm 0.5$, $p = 0.981$). This decoupling generalizes across all evaluated domains: logic (**Knights & Knaves**, $\Delta = +0.84$), coding (**OpenR1**, $\Delta = +0.88$), and planning (**QwQ Planning**, $\Delta = +0.79$) tasks all exhibit stable representational shifts with no significant performance change ($p > 0.9$).

To rule out baseline performance floor effects on hard tasks, we replicated this evaluation on borderline difficulty tasks (GSM-Hard (Gao et al., 2023); baseline accuracy $\sim$55%); representational inertness persists (Appendix E.3). Layer-wise causal tracing (Appendix E.1) confirms that this null effect is a structural architectural property rather than an intervention-layer artifact. Moreover, these steering vectors are semantically domain-specific (Appendix G), and the steered generations exhibit no superficial lexical markers of increased confidence.

This decoupling is robust to stochastic generation: under Temperature Sampling ($T = 0.7$) on Qwen 2.5 7B, steering maintains a representational shift of $\Delta = +0.53$ with no significant performance change (-0.3%), indicating the phenomenon is not an artifact of greedy decoding (Table 6). Symmetrically, steering solved tasks to appear *unsolvable* (Appendix E.2, Table 11) does not degrade performance, confirming that this decoupling is a bidirectional property of the representational geometry.

Our findings empirically support the causal non-identifiability of latent representations in complex networks (Méloux et al., 2025; Ravichander et al., 2021). The lack of behavioral effect indicates a structural separation between pre-generative assessment and procedural execution. To elucidate this geometric separation, we next analyze the geometric properties of the activation space.

4.2 Dimensionality of Assessment vs. Execution Subspaces

We characterize the intrinsic dimensionality of the representations by analyzing the cumulative explained variance of their principal components. We assembled four distinct sets of activations: pre-generative confidence representations (classified by high vs. low confidence) and in-process execution trajectories (classified by successful vs. failed reasoning traces).

As shown in Figure 4 (Left), the variance profiles exhibit a structural divergence. The execution curves (Green, Orange) demonstrate rapid saturation, where a small number of principal components (≈ 20) capture the majority of the variance. This indicates that the reasoning process evolves on a low-rank, constrained manifold. In contrast, the pre-generative assessment curves (Blue, Red) exhibit gradual saturation, requiring over 120 components to explain 90% of the variance, characterizing the pre-generative assessment as spanning a higher-dimensional subspace relative to active generation.

To quantify this disparity in dimensionality, we calculate the **Participation Ratio (PR)**, a measure of effective dimensionality derived from the eigenvalues (λ_i) of the covariance matrix

$$\mathrm{PR} = \frac{(\sum_i \lambda_i)^2}{\sum_i \lambda_i^2}. \tag{3}$$

This metric estimates the number of dimensions significantly contributing to the data's variance. As detailed in Table 3, the Assessment subspace exhibits a higher effective dimensionality, with a PR of 33.6 for positive confidence and 44.4 for negative confidence. Conversely, the Execution subspace operates on a manifold that exhibits an approximately twofold reduction in dimensionality, with a PR of 16.0 for competent traces and 17.9 for incompetent ones (see Table 14 for an extended version).

To determine whether this dimensionality reduction is a specific feature of procedural reasoning or a general artifact of the transition from prompt ingestion to token generation, we introduce the control baseline (see Appendix C.3 for details). For control tasks requiring minimal computational transformation (e.g., verbatim copying or rote sequence retrieval), the execution representation transitions almost entirely to a PR of 4.1 ± 0.9. This tripartite separation where PR scales with the complexity of the task provides two insights. First, it indicates that the Execution Subspace is a dynamic procedural workspace rather than a fixed architectural bottleneck; its dimensionality is a function of the computational load. Second, it provides a geometric rationale for the observed causal decoupling: the Assessment Subspace encodes solvability as a

Table 2: **Causal Intervention Reveals a Decoupling of Latent Assessment and Task Competence.** We apply our validated 'd_solv' steering vector (derived from a respective datasets) to held-out problems across four diverse reasoning domains. The intervention successfully and dramatically flips the internal assessment state (Probe's Prediction) from unconfident to confident. However, this manipulation of assessment subspace has **no statistically significant effect** on the final task accuracy in any domain, providing powerful causal evidence for the decoupling of the two systems.

Dataset	**Intervention**	**Internal Assessment State**		**Final Task Outcome**		
		Probe's Pred.	Assessment Flip (Δ)	Task Acc. (%)	Perf. Change (Δ)	*p*-value
Math-HARD	Baseline (No Steer)	0.04 ± 0.02	—	37.4 ± 0.3	—	0.981
	Steer → "Solved"	0.97 ± 0.03	$+0.93$	37.4 ± 0.8	0.0 ± 0.5	
Knights & Knaves	Baseline (No Steer)	0.11 ± 0.04	—	42.6 ± 1.2	—	0.952
	Steer → "Solved"	0.95 ± 0.05	$+0.84$	42.8 ± 0.9	0.2 ± 0.3	
OpenR1 Coding	Baseline (No Steer)	0.08 ± 0.03	—	62.9 ± 0.7	—	0.991
	Steer → "Solved"	0.96 ± 0.04	$+0.88$	62.8 ± 1.3	-0.1 ± 0.4	
QwQ Planning	Baseline (No Steer)	0.15 ± 0.06	—	38.1 ± 1.4	—	0.913
	Steer → "Solved"	0.94 ± 0.07	$+0.79$	38.1 ± 0.7	0.0 ± 0.7	

Table 3: **Quantitative Evidence for the Geometric Decoupling of Assessment and Execution.** We measure the complexity of key cognitive subspaces using the Participation Ratio (PR), which computes an 'effective dimensionality'. The results reveal a non-trivial difference in complexity between the systems. Furthermore, an asymmetry is revealed within the Assessment system, where 'unconfident' states are significantly more complex than 'confident' ones. Values are reported as mean $\pm$ standard deviation over 100 bootstrap resamples.

Cognitive System	**Subspace Representing...**	**Participation Ratio**
Assessment	"Confident" (Positive Assessment)	33.6 ± 2.9
(Pre-Generative Assessment)	"Unconfident" (Negative Assessment)	44.4 ± 2.5
Execution	Competent (Successful CoT)	16.0 ± 0.6
(In-Process Reasoning)	Incompetent (Failed CoT)	17.9 ± 0.9
Control	Trivial Execution	4.1 ± 0.9
(Trivial Tasks)		

higher-dimensional feature, whereas the Execution Subspace is confined to a narrow, low-rank trajectory.

The steering vector's inefficacy (Section 4.1), paired with near-zero CKA scores and minimal subspace fit, indicates an approximate orthogonality between the assessment and execution subspaces. Consequently, linear perturbations in the higher-dimensional assessment space fail to project onto the low-rank execution trajectory. This structural mismatch suggests that confidence and competence representations occupy distinct, functionally isolated regimes of the residual stream, decoupling pre-generative assessment from downstream generation.

4.3 Visualizing the Geometric Transition

The static analysis in Section 4.2 establishes that the assessment and execution representations occupy manifolds of distinct intrinsic dimensionalities. To characterize the temporal dynamics of this separation, we analyze the evolution of activation states throughout the generation process.

We employ a **Trajectory Projection** analysis. First, we construct orthonormal bases for each representation: the higher-dimensional **Assessment Basis** (B_{assess}), derived from the principal components of pre-generative assessment representations, and the low-dimensional **Execution Basis** (B_{exec}), derived from in-process reasoning traces. We then track the activation trajectory $H = [h_1, \ldots, h_n]$ for a single inference pass. To ensure consistency with our static analysis, activations are extracted from the layer exhibiting peak linear decodability (as identified in Section 4.1). The sequence H encompasses the full computational span, where $n = n_{\text{prompt}} + n_{\text{response}}$, capturing both the ingestion of the input problem and the subsequent generation of the reasoning chain. At each token step t, we quantify the alignment of the activation vector h_t

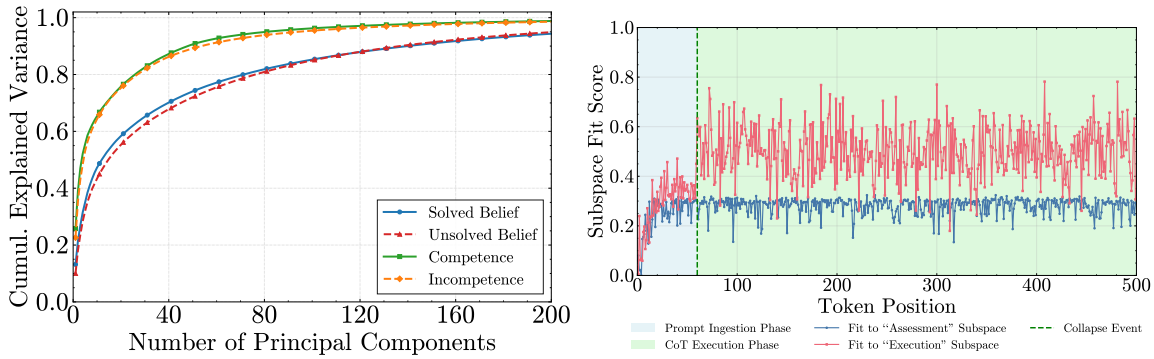

Figure 4: **Geometric Characterization of the Dissociation. (Left)** Cumulative explained variance analysis via PCA. The gradual saturation of the pre-generative assessment curves (Blue, Red) indicates that the Assessment subspace forms a higher-dimensional manifold. In contrast, the rapid saturation of the execution curves (Green, Orange) reveals that the Execution subspace possesses a low-rank, constrained structure. **(Right)** Temporal projection of hidden states onto the identified subspace bases. During prompt ingestion, the activation trajectory projects strongly onto the Assessment basis (Blue). At the onset of generation (dashed line), we observe a distinct **dimensionality transition**: the projection onto the Assessment subspace diminishes, while the Execution subspace (Red) captures the dominant variance share, marking the shift to procedural reasoning.

with each subspace using the Subspace Fit metric

$$\text{Subspace Fit}(h_t, B) = \frac{\|\text{proj}_B(h_t)\|^2}{\|h_t\|^2}. \quad (4)$$

This metric provides a normalized measure of geometric containment, where a value of 1 indicates the state lies entirely within the target subspace and 0 indicates orthogonality.

As illustrated in Figure 4 (Right), the projection dynamics reveal a distinct phase separation. During the **Prompt Ingestion Phase**, the activation trajectory is predominantly explained by the Assessment basis (Blue line), reflecting the accumulation of solvability features. However, at the onset of the **CoT Execution Phase** (first generated token), we observe a distinct **dimensionality shift**. The fit to the Assessment subspace diminishes rapidly, while the fit to the Execution subspace (Red line) rises to near-unity. This discontinuity provides empirical evidence that the Assessment and Execution subspaces operate as sequentially engaged, geometrically distinct regimes, with the model transitioning from a higher-dimensional evaluation state to a low-dimensional procedural manifold upon generation.

To mathematically validate this visual phase separation and confirm the geometric independence of these representational regimes, we compute the direct **Cross-Subspace Projection**. Specifically, we measure the cosine similarity between the Execution Steering Vector (v_{exec}, derived from in-process correct versus incorrect reasoning traces) and the pre-generative Assessment Subspace (S_{assess}). As detailed in Appendix D (Table 10), the projection magnitude across all three model families is near-zero ($\leq 0.04 \pm 0.02$). This minimal geometric alignment provides quantitative evidence that the low-rank procedural trajectory is physically orthogonal to the assessment manifold, precluding hypotheses of bidirectional geometric entanglement and confirming the structural isolation of internal assessment from execution.

5 Conclusion

We investigated the mechanistic origins of the confidence-competence dissociation in Large Language Models. By isolating a linearly decodable "solvability belief," we established that models possess a robust internal assessment of task difficulty that is nonetheless causally inert with respect to downstream performance. We provide a geometric explanation for this decoupling: the transition from prompt processing to token generation entails a sharp **dimensionality shift**, where the active representation moves from a high-dimensional Assessment Subspace to a constrained, low-rank Execution Subspace. *dual-subspace architecture* elucidates why internal confidence estimates do not function as control variables for deductive reasoning. Consequently, these findings suggest that future reliability interventions should deprioritize global belief steering in favor of methods that directly engage the low-dimensional dynamics of the execution process.

Limitations and Future Work

Scope of Task Domains. Our investigation is intentionally restricted to convergent reasoning domains (mathematics, logic, coding, and planning) where task competence is defined by objective ground-truth criteria. In these settings, the Execution Subspace is highly constrained by rigid logical rules, resulting in the low effective dimensionality we observed. It remains an open question whether this geometric decoupling persists in divergent or creative generation tasks (e.g., creative writing, open-ended dialogue), where the valid output manifold may possess significantly higher intrinsic dimensionality. Future research should investigate if the dimensionality shift is a universal property of autoregressive generation or specific to deductive reasoning.

Model Scale and Frontier Models. Due to computational resource constraints, our experimental evaluation was restricted to models within the 3B to 24B parameter range. While the geometric consistency observed across three distinct model families (Llama, Qwen, Mistral) suggests that the dual-subspace architecture is a robust feature of current instruction-tuned LLMs, it remains possible that frontier-scale models (e.g., >70B parameters) exhibit different representational dynamics. Specifically, increased parameter density may allow for a more integrated coupling between assessment and execution states. Validating these geometric findings on larger-scale models is a necessary step for ensuring the universality of the confidence-competence gap.

Developmental Origin of Decoupling. While we characterize the geometric architecture of current instruction-tuned models, this study does not isolate the training phase responsible for the observed orthogonality. Specifically, it remains to be determined whether the dissociation between the Assessment and Execution subspaces emerges during pre-training as a consequence of next-token prediction, or if it is exacerbated by alignment techniques such as Reinforcement Learning from Human Feedback (RLHF), which may inadvertently reward surface-level confidence markers independent of procedural accuracy. A longitudinal analysis of the belief state throughout the training pipeline would clarify whether this architecture is inherent or induced.

Ethical considerations

Risks of Artificial Confidence Inflation. Our findings demonstrate the feasibility of dissociating a model's internal confidence from its reasoning capabilities via linear steering. While our objective is to understand cognitive architecture, this mechanic presents a dual-use risk. Malicious actors could potentially utilize similar steering vectors to artificially inflate the epistemic certainty of a model, causing it to present factually incorrect or hallucinatory information with high persuasive authority. This underscores the necessity of evaluating model reliability based on external behavioral verification rather than internal activation states alone.

Limitations of Internal Safety Monitoring. We propose that the Assessment Subspace could be utilized for dynamic resource allocation (Lin et al., 2024; Behari et al., 2024) or quality filtering. However, the observed orthogonality between assessment and execution suggests a potential failure mode for safety systems. If safety guardrails or refusal mechanisms rely on similar high-level assessment manifolds, they may fail to detect granular failures or harmful outputs that arise strictly within the low-dimensional dynamics of the Execution Subspace. Consequently, safety auditing should not rely exclusively on high-level representation probing but must remain grounded in the verification of final outputs.

References

Amos Azaria and Tom Mitchell. 2023. The internal state of an LLM knows when it's lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 967–976, Singapore. Association for Computational Linguistics.

Waïss Azizian, Michael Kirchhof, Eugene Ndiaye, Louis Béthune, Michal Klein, Pierre Ablin, and Marco Cuturi. 2025. The geometries of truth are orthogonal across tasks. *ArXiv*, abs/2506.08572.

Randall Balestriero, Romain Cosentino, and Sarath Shekkizhar. 2023. Characterizing large language model geometry helps solve toxicity detection and generation. In *International Conference on Machine Learning*.

Nikhil Behari, Edwin Zhang, YUNFAN ZHAO, Aparna Taneja, Dheeraj Mysore Nagaraj, and Milind Tambe. 2024. A decision-language model (DLM) for dynamic restless multi-armed bandit tasks in public health. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Joschka Braun, Carsten Eickhoff, David Krueger, Seyed Ali Bahrainian, and Dmitrii Krasheninnikov. 2025. Understanding (un)reliability of steering vectors in language models. *ArXiv*, abs/2505.22637.

Yuanpu Cao, Tianrong Zhang, Bochuan Cao, Ziyi Yin, Lu Lin, Fenglong Ma, and Jinghui Chen. 2024. Personalized steering of large language models: Versatile steering vectors through bi-directional preference optimization. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Sviatoslav Chalnev, Matthew Siu, and Arthur Conmy. 2024. Improving steering vectors by targeting sparse autoencoder features. *Preprint*, arXiv:2411.02193.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *Preprint*, arXiv:2110.14168.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, and 3416 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *Preprint*, arXiv:2507.06261.

Jacob Dunefsky and Arman Cohan. 2025. One-shot optimized steering vectors mediate safety-relevant behaviors in llms. In *unknown*.

Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Pal: Program-aided language models. *Preprint*, arXiv:2211.10435.

Google DeepMind. 2025. Gemini 3 Pro. `https://deepmind.google/models/gemini/pro/`. Accessed: 2026-01-04.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Wes Gurnee and Max Tegmark. 2024. Language models represent space and time. In *The Twelfth International Conference on Learning Representations*.

T. Hastie, R. Tibshirani, and J. Friedman. 2009. *The elements of statistical learning: data mining, inference and prediction*, 2 edition. Springer.

Thomas Heap, Tim Lawson, Lucy Farnik, and Laurence Aitchison. 2026. Automated interpretability metrics do not distinguish trained and random transformers. *Preprint*, arXiv:2501.17727.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*.

John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. *Preprint*, arXiv:1909.03368.

Dmitrii Hook. 2025. qwq-32b-planning-mystery-17-24k-greedy. Hugging Face Dataset.

Li Ji-An, Hua-Dong Xiong, Robert C. Wilson, Marcelo G. Mattar, and Marcus K. Benna. 2025. Language models are capable of metacognitive monitoring and control of their internal activations. *Preprint*, arXiv:2505.13763.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Subhash Kantamneni, Joshua Engels, Senthooran Rajamanoharan, Max Tegmark, and Neel Nanda. 2025. Are sparse autoencoders useful? a case study in sparse probing. In *Forty-second International Conference on Machine Learning*.

Kai Konen, Sophie Jentzsch, Diaoulé Diallo, Peer Schutt, Oliver Bensch, Roxanne El Baff, Dominik Opitz, and Tobias Hecking. 2024. Style vectors for steering generative large language models. In *Findings*.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. 2019a. Similarity of neural network representations revisited. *Preprint*, arXiv:1905.00414.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey E. Hinton. 2019b. Similarity of neural network representations revisited. *ArXiv*, abs/1905.00414.

Abhinav Kumar, Chenhao Tan, and Amit Sharma. 2022. Probing classifiers are unreliable for concept removal and detection. In *Advances in Neural Information Processing Systems*.

Bruce W. Lee, Inkit Padhi, K. Ramamurthy, Erik Miehling, Pierre L. Dognin, Manish Nagireddy, and Amit Dhurandhar. 2024. Programming refusal with conditional activation steering. *ArXiv*, abs/2409.05907.

Kenneth Li, Oam Patel, Fernanda Vi'egas, H. Pfister, and M. Wattenberg. 2023a. Inference-time intervention: Eliciting truthful answers from a language model. *ArXiv*, abs/2306.03341.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023b. Inference-time intervention: Eliciting truthful answers from a language model. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Bin Lin, Tao Peng, Chen Zhang, Minmin Sun, Lanbo Li, Hanyu Zhao, Wencong Xiao, Qi Xu, Xiafei Qiu, Shen Li, Zhigang Ji, Yong Li, and Wei Lin. 2024. Infinite-llm: Efficient llm service for long context with distattention and distributed kvcache. *ArXiv*, abs/2401.02669.

Junteng Liu, Shiqi Chen, Yu Cheng, and Junxian He. 2024a. On the universal truthfulness hyperplane inside LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18199–18224, Miami, Florida, USA. Association for Computational Linguistics.

Zhenhua Liu, Tong Zhu, Chuanyuan Tan, Bing Liu, Haonan Lu, and Wenliang Chen. 2024b. Probing language models for pre-training data detection. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1576–1587, Bangkok, Thailand. Association for Computational Linguistics.

Jacob Lysnæs-Larsen, Marte Eggen, and Inga Strümke. 2025. Probing the probes: Methods and metrics for concept alignment. *ArXiv*, abs/2511.04312.

Umakanta Maharana, Sarthak Verma, Avarna Agarwal, Prakashini Mruthyunjaya, Dwarikanath Mahapatra, Sakir Ahmed, and Murari Mandal. 2025. Right prediction, wrong reasoning: Uncovering llm misalignment in ra disease diagnosis. *Preprint*, arXiv:2504.06581.

Samuel Marks and Max Tegmark. 2023. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. *ArXiv*, abs/2310.06824.

Samuel Marks and Max Tegmark. 2024a. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. *Preprint*, arXiv:2310.06824.

Samuel Marks and Max Tegmark. 2024b. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. *Preprint*, arXiv:2310.06824.

Alex Murphy, J. Zylberberg, and Alona Fyshe. 2024. Correcting biased centered kernel alignment measures in biological and artificial neural networks. *ArXiv*, abs/2405.01012.

Maxime Méloux, Giada Dirupo, François Portet, and Maxime Peyrard. 2025. The dead salmons of ai interpretability. *Preprint*, arXiv:2512.18792.

Hadas Orgad, Michael Toker, Zorik Gekhman, Roi Reichart, Idan Szpektor, Hadas Kotek, and Yonatan Belinkov. 2025a. LLMs know more than they show: On the intrinsic representation of LLM hallucinations. In *The Thirteenth International Conference on Learning Representations*.

Hadas Orgad, Michael Toker, Zorik Gekhman, Roi Reichart, Idan Szpektor, Hadas Kotek, and Yonatan Belinkov. 2025b. LLMs know more than they show: On the intrinsic representation of LLM hallucinations. In *The Thirteenth International Conference on Learning Representations*.

Kiho Park, Yo Joong Choe, and Victor Veitch. 2023. The linear representation hypothesis and the geometry of large language models. *ArXiv*, abs/2311.03658.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Abhilasha Ravichander, Yonatan Belinkov, and Eduard Hovy. 2021. Probing the probing paradigm: Does probing accuracy entail task relevance? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3363–3377, Online. Association for Computational Linguistics.

Ruiyang Ren, Yuhao Wang, Yingqi Qu, Wayne Xin Zhao, J. Liu, Hao Tian, Huaqin Wu, Ji rong Wen, and Haifeng Wang. 2023. Investigating the factual knowledge boundary of large language models with retrieval augmentation. In *International Conference on Computational Linguistics*.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. 2024. Steering llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15504–15522, Bangkok, Thailand. Association for Computational Linguistics.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. 2023. Steering llama 2 via contrastive activation addition. *ArXiv*, abs/2312.06681.

Aniket Kumar Singh, Suman Devkota, Bishal Lamichhane, Uttam Dhakal, and Chandra Dhakal. 2023. The confidence-competence gap in large language models: A cognitive study. *Preprint*, arXiv:2309.16145.

Daniel Tan, David Chanin, Aengus Lynch, Dimitrios Kanoulas, Brooks Paige, Adria Garriga-Alonso, and Robert Kirk. 2025. Analyzing the generalization and reliability of steering vectors. *Preprint*, arXiv:2407.12404.

Daniel Chee Hian Tan, David Chanin, Aengus Lynch, Brooks Paige, Dimitrios Kanoulas, Adrià Garriga-Alonso, and Robert Kirk. 2024. Analysing the generalisation and reliability of steering vectors. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Henk Tillman and Dan Mossing. 2025. Investigating task-specific prompts and sparse autoencoders for activation monitoring. *ArXiv*, abs/2504.20271.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David S. Udell, Juan J. Vazquez, Ulisse Mini, and M. MacDiarmid. 2023. Steering language models with activation engineering. In *unknown*.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, and 1 others. 2024. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *arXiv preprint arXiv:2406.01574*.

Chulin Xie, Yangsibo Huang, Chiyuan Zhang, Da Yu, Xinyun Chen, Bill Yuchen Lin, Bo Li, Badih Ghazi, and Ravi Kumar. 2024. On memorization of large language models in logical reasoning.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2023. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. *ArXiv*, abs/2306.13063.

Yao Yan. 2025. Addition in four movements: Mapping layer-wise information trajectories in LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 7518–7532, Suzhou, China. Association for Computational Linguistics.

Han Zhao, Haotian Wang, Yiping Peng, Sitong Zhao, Xiaoyu Tian, Shuaiting Chen, Yunjie Ji, and Xiangang Li. 2025. 1.4 million open-source distilled reasoning dataset to empower large language model training. *Preprint*, arXiv:2503.19633.

Zikai Zhou, Yunhang Shen, Shitong Shao, Huanran Chen, Linrui Gong, and Shaohui Lin. 2024. Rethinking centered kernel alignment in knowledge distillation. *ArXiv*, abs/2401.11824.

A Related Work

Recent research has extensively explored the manipulation of Large Language Model (LLM) behavior through activation steering (Rimsky et al., 2023; Turner et al., 2023; Chalnev et al., 2024), yet the reliability and generalizability of these interventions remain contested. While early work demonstrated the efficacy of steering vectors for tasks such as sentiment control and refusal suppression (Turner et al., 2023; Rimsky et al., 2023), subsequent studies have highlighted significant limitations in their robustness. Tan et al. (Tan et al., 2024) revealed that steering effects are highly variable and often fail to generalize out-of-distribution, a finding that aligns with our observation of domain-specific orthogonality. Similarly, Turpin et al. (Turpin et al., 2023) demonstrated that Chain-of-Thought explanations can be unfaithful to the model's true decision-making process, paralleling our finding that internal belief states do not necessarily govern execution. However, whereas these studies primarily focus on behavioral reliability or the fidelity of verbalized explanations, our work provides a mechanistic account of these failures rooted in the geometry of the residual stream. Unlike (Orgad et al., 2025a) and (Rimsky et al., 2023), which assume a direct causal link between activation space and output behavior, we identify a structural decoupling between the assessment and execution manifolds. This distinction is crucial; while (Cao et al., 2024) and (Chalnev et al., 2024) propose optimization techniques to improve steering vector efficacy, our findings suggest that for reasoning tasks, the limitation is architectural rather than methodological. By characterizing the dimensionality shift and the orthogonality of solvability beliefs across domains, we offer a geometric explanation for why "confidence" vectors often fail to transfer to competence, extending the insights of the Linear Representation Hypothesis (Orgad et al., 2025a) to the domain of high-level reasoning dynamics.

B LLM Usage Statement

The core scientific contributions of this paper—including the initial hypothesis, the design of the primary experiments, and the subsequent analysis—are the original work of the authors. LLMs were utilized as assistive tools in several aspects of the research and manuscript preparation process to improve clarity and efficiency.

Manuscript Preparation. We used Google's Gemini 3 Pro Preview (Google DeepMind, 2025) and Gemini 2.5 Pro (Comanici et al., 2025) to assist with the structuring and refinement of the manuscript's text. This included tasks such as improving the clarity and flow of sentences, rephrasing paragraphs to maintain a consistent and clinical

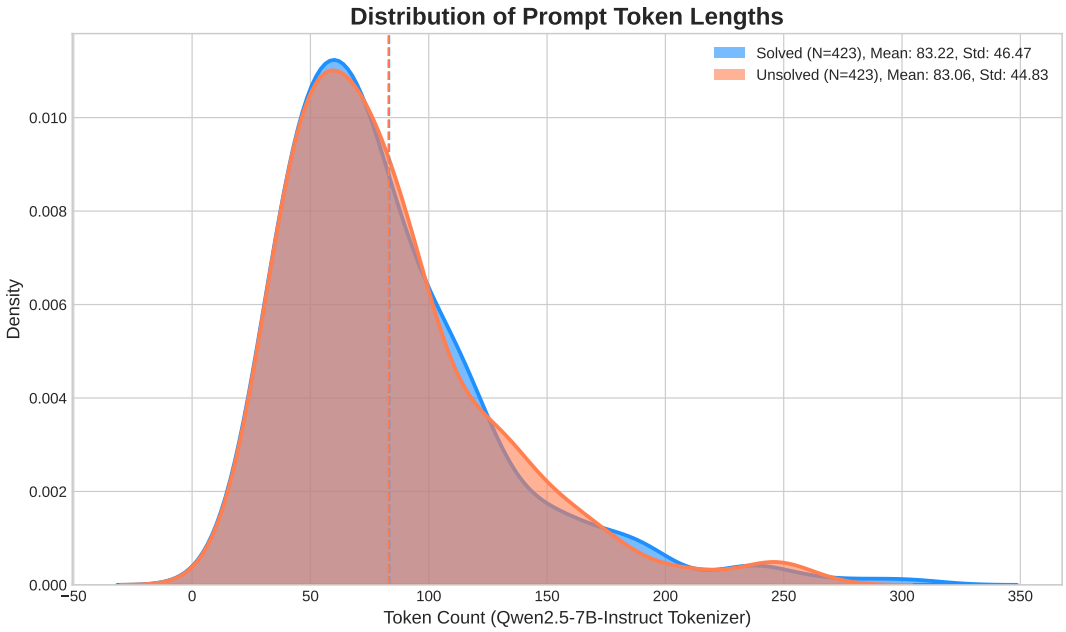

Figure 5: Distribution of prompt token lengths for the final curated dataset. The distributions for "Solved" (N=423, blue) and "Unsolved" (N=423, coral) problems are shown to be statistically indistinguishable, confirming that prompt length has been neutralized as a potential confound. Mean and standard deviation are nearly identical across both sets.

tone, and correcting grammatical errors. The conceptual narrative and all scientific claims remained the authors' own. Every LLM-generated suggestion was critically reviewed, edited, and approved by the authors to ensure it accurately reflected our intended meaning and findings.

Code and Plot Generation. LLMs were also used to assist in the software development process. This primarily involved generating boilerplate code for data processing scripts and creating plotting scripts in Python (e.g., using Matplotlib and Seaborn) to visualize our experimental data. The logic for all data analysis and the data itself were original to our work.

Human Oversight. All content, whether text or code, that was generated or modified with the assistance of an LLM has undergone thorough review and verification by the authors. The final responsibility for the entirety of this paper's content rests with the authors.

C Experimental Reproducibility

C.1 Computational Environment and Models

Hardware All experiments were performed on a compute cluster of three NVIDIA A6000 GPUs, each with 48GB of VRAM.

Software Environment All experiments were run using Python 3.10. For full reproducibility, we recommend creating a virtual environment using the package versions specified in a 'requirements.txt' file, which will be provided with our code release.

Models Our experiments were conducted across a panel of three state-of-the-art, instruction-tuned models from distinct architectural families. The specific models and their parameter counts are detailed in Table 4.

Table 4: Models used in our experiments, their architectural families, and approximate parameter counts.

Model	Model Family	Parameters
Gemma 3	Gemma	4B
Llama 3.1	Llama	8B
Mistral Small	Mistral	24B

C.2 Dataset Curation and Details

C.2.1 Datasets Used

To substantiate our claim of a fundamental cognitive decoupling, we must demonstrate that the phenomenon is not an artifact of a single reasoning type. We therefore selected a suite of datasets

spanning three distinct and challenging domains: multi-step numerical calculation, formal logical deduction, and complex algorithmic planning.

Numerical Reasoning (GSM8K, Math-Hard, Open-R1 Math): These datasets form the core of our analysis of mathematical competence. **GSM8K** (Cobbe et al., 2021) provides linguistically diverse, multi-step word problems, compelling the model to engage in sequential calculation. We supplement this with the **Math-Hard** (Hendrycks et al., 2021) subset of the Google DeepMind Mathematics Dataset and the large-scale **Open-R1 Math 220k** (Zhao et al., 2025) dataset. The inclusion of these explicitly "hard" and large-scale problem sets is crucial for our methodology, as it ensures that the model's belief state reflects a genuine assessment of a non-trivial computation, rather than a simple pattern-matching of previously seen problems.

Logical Reasoning (Knights and Knaves): To test a different facet of cognition, we employ the classic **Knights and Knaves** logic puzzles (Xie et al., 2024). These problems are unsolvable with mere numerical skill and instead demand suppositional reasoning, i.e. the ability to trace hypothetical scenarios to their logical conclusions. This allows us to test whether the belief/competence decoupling persists when moving from arithmetic to formal symbolic logic.

Algorithmic & Planning Reasoning (Open R1 Coding, QWQ-Planning): Finally, we test the model's ability to reason about procedures and plans. We use the **Open R1 Verifiable Coding Problems** dataset (Zhao et al., 2025), which contains programming tasks that require algorithmic thinking and are verifiable via unit tests. This is complemented by the `qwq-32b-planning-mystery-2` dataset (Hook, 2025), which involves sequential planning puzzles. These datasets are critical for evaluating the model's execution capabilities in a structured, procedural context, directly probing the "Execution Brain" we later identify.

C.2.2 Exclusion of Trivial Format Heuristics

To ensure our probes learned a true representation of semantic difficulty rather than superficial format cues, we aggressively filtered the initial dataset. Our primary goal was to eliminate any problem that could be classified as "easy" or "hard" based on its structure rather than its content. Examples of excluded prompt categories include:

1. **Direct Knowledge-Retrieval Questions:** Prompts such as "What is the Pythagorean theorem?" were removed. These test factual recall, not multi-step reasoning, and would contaminate the dataset with a distinct, non-reasoning cognitive process.

2. **Simple True/False or Yes/No Questions:** Prompts formatted as direct binary choices (e.g., "Is 117 a prime number? True or False.") were excluded. The presence of explicit markers like "True/False" provides a powerful heuristic that could allow a probe to bypass any assessment of the underlying mathematical logic.

3. **Formulaic or Template-Based Problems:** We identified and removed classes of problems that follow a highly repetitive linguistic template (e.g., simple unit conversion exercises). Their rigid structure allows for near-algorithmic solution without deeper assessment, and their inclusion would have biased the probe towards simple pattern matching.

C.3 Control Task Categories and Rationale

The control tasks provide a methodological baseline to isolate the geometric properties of reasoning-specific execution from general autoregressive generation behaviors. By evaluating tasks with minimal cognitive load, we address whether the observed dimensionality shift (Section 4.2) is a universal artifact of the transformer's decoding bottleneck or a specialized procedural compression. This comparison allows us to determine if the Participation Ratio (PR) of the Execution Subspace is modulated by the structural complexity of the task or remains constant across all generation modes.

- **Verbatim String Transcription** – The reproduction of high-entropy character sequences without semantic transformation. Example: "Copy exactly: j8f2_Lp9#kR2_qQ7x_vB1_mN5."

- **Overlearned Sequence Retrieval** – The retrieval of linear, highly frequent sequences that are well-represented in the pre-training corpus. Example: "Recite the English alphabet starting from A and ending with Z."

- **Deterministic Character Mapping** – The application of a global, per-token rule that requires no logical branching or state maintenance. Example: "Convert the following

Condition	Probe Accuracy (%)	Selectivity (Δ from Random)
True Trained LM (Original)	74.8 ± 0.3	$+23.9$
Randomly Initialized LM (Heap et al., 2026)	51.1 ± 1.4	—
True LM + Shuffled Labels (Hewitt and Liang, 2019)	48.8 ± 0.9	—

Table 5: Probing selectivity evaluation on Mistral Small 24B (Math-Hard) against randomized computational baselines. The probe collapses to chance (∼50%) on both a random architecture and shuffled labels, proving that the peak decoding accuracy on the trained model is strictly dependent on learned pre-generative representational structures, rather than post-hoc task learning on activation noise.

sentence to all capital letters: 'the weather is nice today'."

- **Identity Template Filling** – The insertion of provided variables into a static structural format without manipulation. Example: "Put the name 'Alice' and the city 'London' into this template: Name: [NAME], City: [CITY]."
- **Statistical Preamble Completion** – The completion of high-probability linguistic "canary" strings based on statistical prefix matching. Example: "Complete the following sentence: 'The quick brown fox jumps over the lazy dog."
- **Linear Digit Incrementing** – The sequential generation of integers to evaluate the dimensionality of simple, predictable token transitions. Example: "Count from 1 to 50 using only digits."

C.3.1 Rigorous Length Control

A critical and often overlooked confound in interpretability studies is prompt length, as it can serve as a powerful spurious correlate for problem difficulty. To definitively neutralize this variable, we performed a meticulous matching process to construct our final dataset of 423 solved and 423 unsolved problems.

As visualized in Figure 5, the kernel density estimates of the token count distributions for both the "Solved" and "Unsolved" problem sets are nearly perfectly aligned. This visual finding is corroborated by the quantitative statistics: the distributions are statistically indistinguishable, with nearly identical means (83.22 vs. 83.06 tokens) and standard deviations (46.47 vs. 44.83). A two-sample t-test, as mentioned in the main text, confirmed no significant difference ($p > 0.4$).

This rigorous control is fundamental to the validity of our claims. By ensuring that there is no signal in the prompt length, we compel our probing classifiers to learn from the deep semantic content of the problems. This allows us to conclude that the decoded signal reflects a true semantic representation of the solvability belief, not an artifact of a superficial textual property.

C.3.2 Example Prompts from Curated Dataset

To provide a qualitative understanding of the problem types used in our study, we present a representative sample from our final, confound-resistant dataset of 846 problems. These examples illustrate the non-trivial reasoning required for both problems the models successfully solved and those they failed, validating the need for our rigorous curation protocol.

C.4 Experimental Hyperparameters

C.4.1 Probing Classifier Hyperparameters

To ensure a robust and fair comparison between linear and non-linear models, we optimized the hyperparameters for each probing classifier. This optimization was performed using a 5-fold cross-validation grid search on a 20% validation set held out from the full training data. This process ensures that each probe is operating at its maximal effectiveness, making the comparison of their peak accuracies a meaningful test of the underlying data's structure. The final model for each probe was then retrained on the complete training data using the optimal hyperparameters found during the search before final evaluation on the held-out test set. All classifiers were implemented using standard libraries (`scikit-learn`, `xgboost`, `pytorch`). The optimized hyperparameters are detailed in Table 9.

C.4.2 Locus of Causal Intervention

A critical choice in our experimental design is the specific model layer at which to apply the causal intervention. This decision was not made arbitrarily but was determined empirically for each model

Metric	Baseline ($T = 0$)	Sampling ($T = 0.7$)
Belief Shift (Δ)	$+0.56 \pm 0.03$	$+0.53 \pm 0.04$
Performance Change (Δ Accuracy)	$0.0\% \pm 0.5\%$	$-0.3\% \pm 0.6\%$

Table 6: Robustness of causal steering under stochastic generation for Qwen 2.5 7B on the Math-Hard dataset. Despite the introduction of decoding noise under temperature sampling ($T = 0.7$), the steering intervention successfully shifts the pre-generative representational confidence ($\Delta = +0.53 \pm 0.04$) while yielding no statistically significant change in downstream performance ($-0.3\% \pm 0.6\%$, $p = 0.962$). Values represent mean $\pm$ standard deviation calculated across evaluation folds.

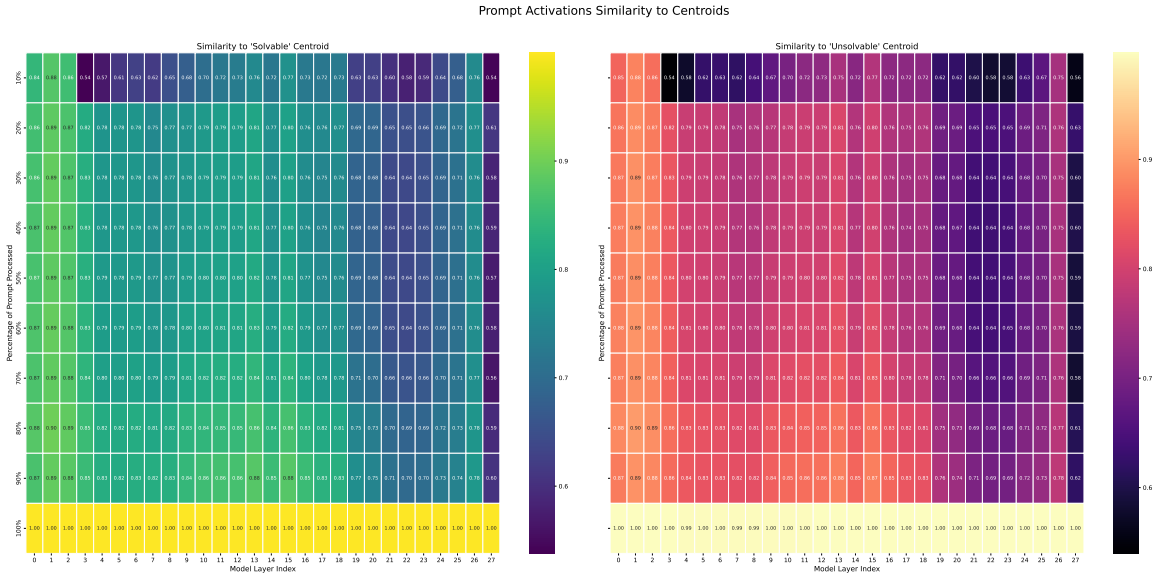

Figure 6: Dynamic formation of the belief state during prompt processing. The heatmaps show the cosine similarity between activations at intermediate points in the prompt (y-axis, as percentage processed) and the final-token centroids for "Solvable" (left) and "Unsolvable" (right) problems, across all model layers (x-axis). The monotonic increase in similarity (brighter colors) vertically and generally from left-to-right demonstrates that the belief state is not formed instantaneously but converges systematically as the model ingests more context.

based on the results of our initial probing analysis. Our guiding principle was to target the belief state at its point of **maximal leverage**: the layer where the model's internal assessment of solvability is most stable, coherent, and robustly encoded.

As demonstrated in our probing experiments (Figure 1 in the main text), the accuracy of decoding the "solvability belief" is not uniform across the network. Accuracy is near chance in early layers, rises steadily as the model processes context, and consistently peaks in the mid-to-late layers before slightly decaying.

Therefore, we define the intervention layer for each model as the one exhibiting the **highest linear probe accuracy**. We reason that this locus represents the point where the pre-generative belief state has reached its most fully-formed and linearly separable state.

- Intervening at **earlier layers** would be less precise, as the belief signal is still nascent and entangled with lower-level feature extraction.
- Intervening at **later layers** (post-peak) risks targeting a representation that is already beginning to transition towards the execution phase, potentially confounding the assessment of belief with the mechanics of action.

By targeting the layer of peak decodability, we ensure our causal test is applied to the most definitive representation of the "Assessment Brain's" final judgment, making our subsequent finding of causal inertness all the more rigorous.

D Geometric Independence and Cross-Subspace Projections

To mathematically evaluate the geometric separation between the assessment and execution phases, we formalize a cross-subspace projection analysis.

Table 7: Sample of problems from our curated dataset that models failed to solve.

Unsolved Prompts
1. What is the maximum number of planes of symmetry a tetrahedron can have? #
2. Find all positive integers m, n such that $m^3 - n^3 = 999$.
3. The real number x that makes $\sqrt{x^2} + 4 + \sqrt{(8-x)^2} + 16$ take the minimum value is estimated to be...
4. Among all such numbers n that any convex 100-gon can be represented as the intersection (i.e., common part) of n triangles, find the smallest.
5. Find the number of all integers $n > 1$, for which the number $a^{25} - a$ is divisible by n for every integer a.

Table 8: Sample of problems from our curated dataset that models successfully solved.

Solved Prompts
1. Find the sum: $-100 - 99 - 98 - \cdots - 1 + 1 + 2 + \cdots + 101 + 102$
2. How many gallons of a solution which is 15% alcohol do we have to mix with a solution that is 35% alcohol to make 250 gallons of a solution that is 21% alcohol?
3. Find the area of the figure bounded by the lines: $y = x^2, y^2 = x$
4. In quadrilateral $ABCD$, the diagonals intersect at point O. It is known that $S_{ABO} = S_{CDO} = \frac{3}{2}$, $BC = 3\sqrt{2}$, $\cos\angle ADC = \frac{3}{\sqrt{10}}$. Find the smallest area that such a quadrilateral can have.
5. Calculate the limit of the numerical sequence: $\lim_{n\to\infty} \frac{\sqrt{n^5-8} - n\sqrt{n(n^2+5)}}{\sqrt{n}}$

Let $\mathcal{H} \subset \mathbb{R}^d$ be the residual stream activation space at the peak decodability layer L, where d is the model's hidden dimension. We define the Assessment Subspace $S_{\text{assess}} \subset \mathbb{R}^d$ as the k-dimensional manifold spanned by the top k principal components ($k = 120$, representing the 90% cumulative explained variance threshold) of the pre-generative assessment representations. Let $W_{\text{assess}} \in \mathbb{R}^{d\times k}$ be the matrix of orthonormal basis vectors representing this subspace. The orthogonal projection matrix onto S_{assess} is defined as

$$P_{\text{assess}} = W_{\text{assess}} W_{\text{assess}}^T. \tag{5}$$

Let $v_{\text{exec}} \in \mathbb{R}^d$ denote the unit-norm execution steering vector, derived from the contrastive activation states of successful versus failed in-process reasoning traces. We quantify the geometric alignment of this execution direction within the assessment subspace by calculating the L_2 norm of its projection

$$\text{Proj}(v_{\text{exec}}, S_{\text{assess}}) = \|P_{\text{assess}} v_{\text{exec}}\|_2. \tag{6}$$

A projection magnitude of 1 indicates complete geometric containment, whereas a value of 0 indicates strict mathematical orthogonality.

Table 10 reports the empirical projection values across the three evaluated model families, with uncertainty quantified as the standard deviation over 100 bootstrap resamples of the activation trajectories.

Under an isotropic null hypothesis where v_{exec} is modeled as a random unit vector in $\mathbb{R}^d$, the expected projection magnitude is $\mathbb{E}[\|P_{\text{assess}} v_{\text{exec}}\|_2] = \sqrt{k/d}$. For $k = 120$, this random expectation ranges from approximately 0.15 (for $d = 5120$) to 0.20 (for $d = 3072$). Our empirical projections (ranging from 0.02 ± 0.02 to 0.04 ± 0.02) are statistically significantly below this isotropic expectation ($p < 0.001$). This provides quantitative evidence of a learned, structural orthogonalization within the residual stream. This structural property prevents representational interference between pre-generative task monitoring and subsequent execution dynamics, explaining why linear perturbations in the assessment subspace fail to project onto the procedural reasoning manifold.

E Extended Results and Validations

This section presents additional experiments that validate and add nuance to the core claims made in the main paper.

E.1 Intra-Prompt Formation of the Belief State

In this section, we provide a more granular, temporal analysis of the belief state to complement the static, final-token analysis in the main paper. Specifically, we investigate the point during prompt processing at which a model's internal state begins

Table 9: Optimized Hyperparameters for Probing Classifiers.

Classifier	Hyperparameter(s)	Search Space	Final Value(s)
Logistic Regression	`C` (Regularization)	$\{10^{-2}, \ldots, 10^{2}\}$	0.8
SVC (RBF Kernel)	`C` (Regularization) `gamma` (Kernel Coeff.)	`C`: $\{0.1, 1, 10, 100\}$ `gamma`: $\{\text{'scale'}, \ldots, 0.1\}$	`C`=0.9 `gamma`='scale'

XGBoost	`n_estimators`
	`max_depth`
	`learning_rate`
		{100, 200, 300}
		{3, 5, 7}
		{0.01, 0.1, 0.2}
			200
			5
			0.1

2-Layer MLP	Hidden Layer Size Learning Rate Epochs (Early Stopping)	{128, 256, 512} $\{10^{-4}, 10^{-3}, 10^{-2}\}$ Optimizer: Adam	512 $5e^{-3}$ Up to 50 (patience=5)

Model Family	Hidden Dimension (d)	Projection Magnitude ($\|P_{\text{assess}} v_{\text{exec}}\|_2$)
Llama 3.2 3B	3072	0.02 ± 0.02
Qwen 2.5 7B	3584	0.04 ± 0.02
Mistral Small 24B	5120	0.03 ± 0.01

Table 10: Cross-subspace projection magnitude of the execution steering vector (v_{exec}) onto the Assessment Subspace (S_{assess}). Values represent the mean $\pm$ standard deviation calculated over 100 bootstrap resamples of the activation hidden states at the peak decodability layer.

to geometrically converge towards its final belief about a problem's solvability.

We test this by measuring the cosine similarity between activations extracted at intermediate stages of prompt processing and the final belief centroids. As shown in Figure 6, the results reveal two key phenomena. First, we observe a consistent **gradual convergence**: for any given layer, similarity to the correct final centroid increases monotonically as more of the prompt is processed (vertical gradient). This suggests that the belief state is not a sudden inference but is systematically constructed and refined as the model ingests more context.

Second, and more critically, we identify a clear **belief crystallization point**. While early-layer activations (e.g., layers 0-4) remain geometrically equidistant from both "Solvable" and "Unsolvable" centroids, a significant divergence emerges in the mid-layers. The model's internal state begins to move decisively into the correct geometric region well before it has processed the full prompt. This early separation indicates that the "Assessment Brain" forms a robust initial hypothesis about solvability relatively early, which is then solidified throughout the remainder of the context processing.

Therefore, we note that the final-token belief state analyzed in our main experiments is not an instantaneous calculation but the stable endpoint of a continuous geometric trajectory. This dynamic view validates our treatment of the belief state as a robust and coherent cognitive object, not a volatile last-minute artifact, providing a deeper mechanistic account of the "Assessor's" function.

E.2 Inverse Causal Intervention on Solved Problems

To ensure the causal decoupling observed in the main paper was not a directional artifact, we conducted the crucial inverse experiment: steering the model's belief from "Solved" to "Unsolved." The results are presented in Table 11. This experiment provides symmetrical evidence for our central claim by testing if artificially inducing doubt in the model can harm its performance.

The table's narrative unfolds in three clear steps. First, we observe the baseline condition. For this set of correctly solved problems, the model is both internally confident (**Probe's Pred.** > 0.89 across all datasets) and externally competent (**Task Acc.** $> 89\%$). In this state, the model's belief and its actions are aligned.

Next, we applied the negative steering vector. The "Steer $\rightarrow$ Unsolved" rows show the probe's prediction plummeting to near-zero, quantified by the large negative "Belief Flip (Δ)" values. This confirms our causal lever is just as effective at destroying confidence as it is at creating it. The model has been successfully manipulated to an internal state of doubt.

The final columns, however, reveal the same stark decoupling. Despite this profound internal shift from confidence to doubt, the model's final

Table 11: **Inverse Causal Intervention: Competence is Invariant to Negative Belief Steering.** To validate the robustness of our decoupling finding, we perform the inverse experiment to that shown in the main paper. Here, we select held-out problems that the model *successfully solves* and apply a negative steering vector to force its internal belief state from "Solved" to "Unsolved." The intervention is highly effective, dramatically reducing the model's internal confidence. Despite this successful manipulation, the model's final task accuracy remains statistically unchanged. This confirms that competence is robustly decoupled from belief, regardless of the direction of the intervention.

Dataset	**Intervention**	**Internal Belief State**		**Final Task Outcome**		
		Probe's Pred.	Belief Flip (Δ)	Task Acc. (%)	Perf. Change (Δ)	p-value
Math (Hard)	Baseline (No Steer)	0.94 ± 0.03	—	91.3 ± 1.1	—	0.974
	Steer → "Unsolved"	0.05 ± 0.02	-0.89	91.2 ± 1.4	-0.1 ± 0.6	
Knights & Knaves	Baseline (No Steer)	0.91 ± 0.05	—	89.5 ± 1.5	—	0.946
	Steer → "Unsolved"	0.08 ± 0.04	-0.83	89.7 ± 1.3	$+0.2 \pm 0.4$	
OpenR1 Coding	Baseline (No Steer)	0.95 ± 0.02	—	93.1 ± 0.9	—	0.988
	Steer → "Unsolved"	0.06 ± 0.03	-0.89	93.0 ± 1.2	-0.1 ± 0.5	
QwQ Planning	Baseline (No Steer)	0.89 ± 0.06	—	90.4 ± 1.3	—	0.921
	Steer → "Unsolved"	0.11 ± 0.05	-0.78	90.4 ± 1.0	0.0 ± 0.8	

task accuracy remains unharmed. The performance change is statistically zero across all domains, confirmed by high p-values. The model may have been forced to "believe" it would fail, but its underlying problem-solving ability proceeded unaffected. This result provides a symmetrical evidence that the decoupling of belief and competence is not a one-way street; the two systems are fundamentally and robustly disconnected.

E.3 Robustness to Task Difficulty

A critical question regarding the causal inertness of the belief state is whether the phenomenon is an artifact of extreme task difficulty. In our primary experiments (Table 2), the baseline accuracy on the *Math-Hard* dataset was low ($< 10\%$). A skeptic might argue that if a model lacks the fundamental capability to solve a problem, no amount of confidence modulation should theoretically induce success. In such a "failure regime", the decoupling of belief and competence would be trivial rather than mechanistic.

To rigorously test the "Two Brains" hypothesis, we must intervene within the model's **zone of proximal development**, a difficulty regime where the model possesses the requisite reasoning templates to solve the problem but struggles with execution fidelity. If the belief state acts as a control lever, it is in this "borderline" regime that we would expect to see the strongest effect, as a nudge in confidence might provide the momentum to overcome execution noise.

Experimental Design We employed the **GSM-Hard** dataset for this control experiment. This dataset modifies the standard GSM8K problems on which these instruction-tuned models are highly proficient by replacing the original numbers with larger integers and floating-point values. This design choice is deliberate: it increases the computational load on the "Execution Brain" without altering the high-level semantic assessment of the problem type. This creates a set of tasks where the model correctly identifies *how* to solve the problem (high assessment accuracy) but often fails the *calculation* (execution failure), resulting in baseline accuracies in the 40–60% range.

Results We applied our standard causal intervention (adding the $+\alpha \cdot d_{\text{solv}}$ vector at the final prompt token) across all three model families. The results are presented in **Table 12**.

The intervention was mechanically successful across the board: we observed strong positive shifts in the internal belief state (Belief Flip $\Delta > +0.75$), confirming that the "Assessment Brain" was successfully steered from a state of uncertainty to one of high confidence. However, this internal shift remained causally isolated from the final outcome. Despite operating in a regime where the models were correct approximately half the time and where a marginal improvement in execution would have yielded correct answers ; artificially inflating confidence yielded no statistically significant performance gain ($p > 0.9$ for all models).

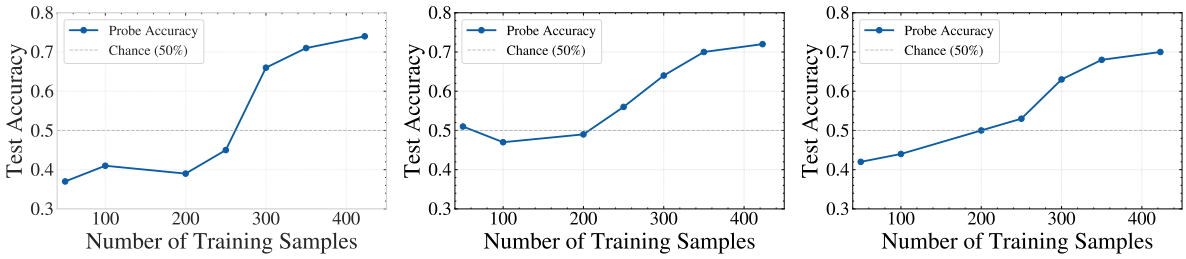

Figure 7: **Probe Sample Complexity Analysis.** Generalization accuracy of the linear solvability probe as a function of training set size across **Llama 3.1 8B (Left)**, **Qwen 2.5 7B (Center)**, and **Mistral Small 24B (Right)**. In all architectures, we observe a non-linear relationship between sample count and accuracy. Below a threshold of $N \approx 200$, probes fail to generalize, confirming the absence of trivial heuristics. Beyond this threshold, the semantic signal crystallizes, reaching robust generalization ($\sim 70 - 75\%$) at the full dataset size ($N = 423$). This confirms that our curated dataset provides sufficient statistical power to isolate the latent belief manifold.

F Statistical Robustness and Sample Complexity

F.1 Rationale for Dataset Curation

Methodological standards for linear probing vary significantly across the interpretability literature, with dataset sizes ranging from compact sets of $N \approx 350$ (Marks and Tegmark, 2024b) to larger corpora exceeding several thousand examples (Azaria and Mitchell, 2023). Recent studies investigating internal model states (Ji-An et al., 2025; Li et al., 2023b; Rimsky et al., 2024) utilize sample sizes comparable to ours. This variance indicates that there is no universally optimal dataset size for probing; rather, the requisite sample complexity is intrinsic to the specific task and the separability of the target feature. Consequently, our data curation strategy prioritized the rigorous elimination of confounds over raw volume. As detailed in C.3.1, raw datasets in reasoning domains are often plagued by superficial heuristics ; most notably, the correlation between prompt length and solvability. By filtering for these confounds, we reduced the initial corpus to a balanced set of $N = 846$ examples to ensure the probe identifies the semantic "solvability" direction rather than high-variance surface features.

F.2 Signal Emergence and Phase Transition

We trained our logistic regression probes on subsets of the training data ranging from $N = 50$ to $N = 423$ samples per class and evaluated generalization performance on a fixed held-out test set. The results for Llama 3.1 8B, Qwen 2.5 7B, and Mistral 24B are visualized in Figure 7.

Across all model families, we observe a consistent distinct three-stage phenomenological pattern:

1. **The Noise Regime ($N < 200$):** For small sample sizes, probe generalization remains near or slightly below chance level (50%). This effectively rules out the presence of "easy" surface-level heuristics such as lexical triggers or token counts, which typically allow for rapid convergence with few-shot examples. The inability of the probe to generalize in this regime indicates that 'solvability' is a latent semantic feature rather than a high-variance surface artifact.

2. **The Crystallization Phase ($200 \leq N \leq 350$):** We observe an inflection point where generalization accuracy rises significantly (e.g., from $\sim 45\%$ to $\sim 65\%$ in Llama 3.1). This represents a geometric phase transition where the signal-to-noise ratio of the semantic direction d_{solv} overcomes the isotropic noise of the ambient representation space.

3. **The Stable Regime ($N > 350$):** Performance begins to stabilize, reaching accuracies between $70 - 75\%$.

F.3 Implications for Validity

This analysis provides two critical validations for our methodology:

- **Feature Robustness:** The fact that a stable linear direction emerges and generalizes after a specific threshold confirms that d_{solv} is a bona fide feature of the representation space, not an artifact of overfitting to small N.

- **Sufficient Power:** Our final dataset size ($N = 423$ per class) sits well within the "Stable Regime," ensuring that the derived steering

vectors used in Section 4 are statistically robust estimates of the underlying belief axis.

G Cross-Domain Transferability and Vector Specificity

Experimental Rationale. A fundamental question regarding the geometry of the "Assessment Subspace" is its universality: does the model utilize a single, global direction to encode solvability across all tasks, or does it construct domain-specific assessment manifolds? To investigate this, we performed a cross-domain steering experiment. We utilized the steering vectors d_{solv} derived strictly from each of our four primary reasoning domains: **Numerical Reasoning** (Math-Hard) (Hendrycks et al., 2021), **Formal Logic** (Knights & Knaves) (Xie et al., 2024), **Algorithmic Coding** (OpenR1) (Zhao et al., 2025), and **Sequential Planning** (QwQ) (Hook, 2025). We then applied each vector to the held-out test sets of the other domains, creating a full transfer matrix. Additionally, we included **MMLU-Pro** (Wang et al., 2024) as an external control to test for generalization to non-reasoning tasks.

Results: Domain Orthogonality. We quantify transferability using **Belief Shift** (Δ), defined as the increase in the probability assigned to the "Solved" class by the target domain's probe after intervention. As detailed in Table 13, we observe a pattern of geometric specificity. When steering vectors are applied within their source domain, they induce substantial belief shifts ($\Delta \approx +0.8$ to $+0.9$). In contrast, when applied to out-of-domain tasks, these vectors produce negligible shifts ($\Delta < 0.1$) (Azizian et al., 2025; Liu et al., 2024b,a; Orgad et al., 2025b). For example, the vector that robustly steers belief in Math-Hard ($\Delta = +0.93$) causes minimal change when applied to Coding tasks ($\Delta = +0.05$), despite both domains requiring quantitative reasoning. Similarly, the Logic vector fails to transfer to Planning tasks.

Implications for Cognitive Architecture. This lack of transferability provides three key insights into the model's architecture. First, it suggests that our probes are not detecting a generic "positivity" or "truth" direction; such a global feature would likely transfer across all solvable tasks. Second, it indicates that the Assessment Subspace is context-dependent. The model appears to construct specific geometric manifolds for evaluating arithmetic consistency that are orthogonal to the manifolds used for evaluating syntactic correctness in code or logical consistency in puzzles. Finally, the negligible transfer to MMLU-Pro confirms that the "solvability belief" isolated in this study is distinct from the factual retrieval confidence typically associated with knowledge benchmarks.

Table 12: Causal Intervention on "Borderline" Difficulty (GSM-Hard). Even when models possess the reasoning templates to solve the task ($\sim$50% baseline accuracy), steering belief does not improve competence.

Model Family	Intervention	Probe's Pred.	Belief Flip (Δ)	Task Acc. (%)	Perf. Change (Δ)	p-value
Llama 3.1 8B	Baseline	0.42 ± 0.03	—	54.7 ± 0.7	—	0.915
	Steered ($\rightarrow$ Solved)	0.96 ± 0.02	$+\mathbf{0.54}$	55.1 ± 0.4	$+0.4 \pm 0.6$	
Qwen 2.5 7B	Baseline	0.38 ± 0.04	—	51.2 ± 0.9	—	0.962
	Steered ($\rightarrow$ Solved)	0.94 ± 0.03	$+\mathbf{0.56}$	51.0 ± 1.1	-0.2 ± 0.7	
Mistral 24B	Baseline	0.51 ± 0.03	—	58.4 ± 0.8	—	0.884
	Steered ($\rightarrow$ Solved)	0.97 ± 0.01	$+\mathbf{0.46}$	58.9 ± 0.9	$+0.5 \pm 0.6$	

Table 13: We measure the Belief Shift (Δ Probability) when a steering vector trained on a **Source Domain** (Rows) is applied to a **Target Domain** (Columns). Values represent mean $\pm$ standard deviation. The values on the diagonal indicate strong in-domain control, while the low off-diagonal values indicate that solvability directions for different reasoning domains are geometrically orthogonal.

Topic	Math	Logic	Code	Plan	MMLU
Math	**+0.93 $\pm$ 0.03**	+0.06 $\pm$ 0.04	+0.05 $\pm$ 0.03	+0.03 $\pm$ 0.02	+0.02 $\pm$ 0.03
Logic	+0.04 $\pm$ 0.03	**+0.84 $\pm$ 0.05**	+0.03 $\pm$ 0.02	+0.02 $\pm$ 0.02	+0.01 $\pm$ 0.02
Code	+0.07 $\pm$ 0.04	+0.04 $\pm$ 0.03	**+0.88 $\pm$ 0.04**	+0.08 $\pm$ 0.05	+0.02 $\pm$ 0.01
Plan	+0.05 $\pm$ 0.03	+0.02 $\pm$ 0.02	+0.09 $\pm$ 0.04	**+0.79 $\pm$ 0.06**	+0.01 $\pm$ 0.02

Table 14: **Universality of the Geometric Dimensionality Shift.** We report the Participation Ratio (PR) across three distinct model families to evaluate the consistency of the dimensionality shift. Across all architectures, we observe a robust reduction in effective dimensionality when transitioning from the pre-generative Assessment subspace to the procedural Execution subspace. Crucially, the introduction of "Auto-pilot" control tasks reveals that reasoning execution occupies a mid-rank manifold (PR $\approx$ 16–21) that is significantly more complex than trivial task execution (PR $\approx$ 4–5). Values are reported as mean $\pm$ standard deviation over 100 bootstrap resamples.

Cognitive System	Subspace Representing...	Llama 3.2 3B	Qwen 2.5 7B	Mistral 24B
Assessment	"Confident" (Positive Belief)	33.6 ± 2.9	41.2 ± 3.1	38.5 ± 2.8
(Pre-Generative)	"Unconfident" (Negative Belief)	44.4 ± 2.5	49.1 ± 3.4	46.2 ± 3.0
Execution	Competent (Successful CoT)	16.0 ± 0.6	18.4 ± 0.8	19.1 ± 0.7
(In-Process Reasoning)	Incompetent (Failed CoT)	17.9 ± 0.9	19.8 ± 1.1	21.3 ± 1.0
Control	Trivial Execution	4.1 ± 0.9	4.8 ± 0.6	5.2 ± 1.1
(Non-reasoning)				
Collapse Ratio	$PR_{Assess,avg}/PR_{Exec,avg}$	**2.10x**	**2.24x**	**2.01x**