STRATEGIC DIVERSITY DOES NOT SCALE: BOUNDED GEOMETRY OF LLM COLLECTIVE BEHAVIOR

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-model AI systems are often motivated by an expectation that independently developed language models provide complementary judgments. We examine this expectation using 50 LLMs from 26 model families evaluated on 20 strategic-game conditions spanning signaling, coordination, Bayesian reasoning, and deception. We represent model responses using text embeddings and study the geometry of both individual responses and model-level behavioral signatures. As the evaluated ecosystem grows, we observe that dimensionality of the pooled response space increases with sharply diminishing marginal returns. At 50 models, principal component analysis (PCA) over model-level signatures requires 16 components to explain 90% of between-model variance. Model signatures also exhibit high pairwise similarity and substantial cross-family geometric alignment. These patterns persist in action-only representations, alternative dimensionality estimators, and an open-ended generation dataset. Together, our results indicate that nominal growth in models and providers produces less behavioral expansion than model count alone would suggest. We further investigate strategic diversity when models are aggregated following three patterns: majority vote, multi-agent discussion, cross provider composition. Majority voting reinforces model's dominant responses. Multi-agent discussion among diverse models produces no clear gain in strategic score, while critique from another provider often changes the decision but does not reliably improve it. Effective collective systems thus require measured coverage of useful behaviors, not simply more models, providers, or calls.

1 INTRODUCTION

Modern AI workflows increasingly select among or combine multiple large language models (LLMs) through model routing (Ong et al., 2025), aggregation (Jiang et al., 2023), cross-provider composition, and multi-agent discussion (Wu et al., 2024). Such designs are motivated by an implicit assumption: if models are built by different organizations, trained on different data, and with different post-training pipelines, deployed with different architectures, then their judgments should be meaningfully diverse. When one model fails, another should fail differently.

As multi-model workflows emerge as advisors and collaborators in high-stakes decision-making, it is critical that they provide different perspectives on the same strategic situation rather than merely different surface phrasings. In this paper, we focus on *strategic reasoning tasks*, where quality and correctness depend not only on producing a fluent answer, but also on anticipating how other actors will behave. Such tasks arise in negotiation, safety evaluation, persuasion, tool use, and market design, where an AI system must predict how another actor will respond.

We test how strategic behavioral diversity changes as the evaluated LLM ecosystem grows. We call models *strategically diverse* when they produce systematically different policies in response to the same classes of incentives and information structures. We evaluate 50 LLMs from 26 model families on 20 strategic-game conditions spanning signaling, coordination, Bayesian reasoning, and deception. Each model is evaluated repeatedly, producing 9,871 valid responses.

We consider two forms of variation. *Response-space variation* includes differences across models, games, instances, and repeated samples. *Between-model variation* is measured using behavioral signatures that average each model's responses within games and then across the game suite.

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.

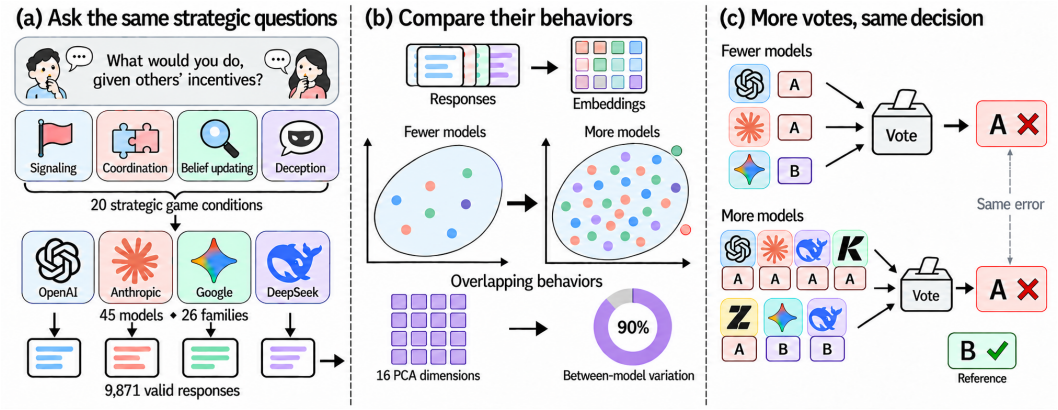

Figure 1: **Overview.** We collect responses from 50 models on a common strategic-game suite and embed the responses in a shared representation space. We measure total response-space expansion using balanced PCA and isolate persistent between-model variation using model-level behavioral signatures. Across these complementary analyses, adding models produces diminishing geometric expansion, and models from different families occupy substantially overlapping behavioral regions.

Our results consistently indicate diminishing behavioral expansion. Pooled-response PCA grows from 76.2 at five models to 137.0 at 50 models, but its marginal rate of growth declines substantially. The between-model analysis is more compact: PCA over the 50 model-level signatures yields 16 dimensions. Across adequately expressive autoencoder configurations, reconstruction gains become small between 18 and 24 bottleneck dimensions, although the exact plateau depends on the architecture. We refer to the resulting concentration of nominally distinct models in an overlapping behavioral geometry as *geometric monoculture*.

We test the practical consequence of this geometry with three common ways of combining model outputs. Majority voting raises average strategic score by making a model's most common action more likely to determine the final answer. Discussion among diverse models produces responses that move in new geometric directions, but those directions do not yield a clear improvement in strategic score. Cross provider critique changes the proposed action in 40.2% of transcripts, yet improves the score in only 4.6%. These results separate behavioral movement from strategic progress: combining models can make answers more consistent or more varied without adding the missing useful behavior.

Related work. Prior work evaluates LLMs in strategic interactions on a per-model basis (Gandhi et al., 2023; Brookins & DeBacker, 2023; Guo, 2023; Fan et al., 2024; Akata et al., 2025); we provide a geometric embedding space that enables ecosystem-level comparison. Weight-space model merging (Wortsman et al., 2022; Yadav et al., 2023; Yu et al., 2024) and activation-space task vectors (Ilharco et al., 2023; Zou et al., 2023) require access to model internals; our approach operates entirely from API-level text outputs. The dimensional hierarchy we observe parallels findings in neuroscience, where cognitive dimensionality scales with task complexity (Cunningham & Yu, 2014; Jazayeri & Ostojic, 2021), and complements the superposition hypothesis in mechanistic interpretability (Elhage et al., 2022; Templeton et al., 2026) by showing that *behavioral* outputs collapse to far fewer dimensions than internal representations.

2 Evaluation Suite and Data Collection

Goal. Our goal is to test whether independently developed models exhibit diverse behaviors in strategic reasoning tasks.

LLMs. We collect responses from 50 LLMs spanning 26 independent families. Models we consider include both open-weight and closed-source ones with different architectures and scales. More details are in Appendix B. We fix temperature at 0.7 to standardize decoding across providers. We fall back to provider defaults when temperature control is unavailable.

Strategic Reasoning Tasks. We construct 20 strategic games across 4 categories: 5 **signaling**, 5 **coordination**, 5 **Bayesian reasoning**, and 5 **deception** games. Each game has explicit payoffs, unambiguous action spaces, and well-defined outcomes for downstream analysis. See Appendix A for more detailed descriptions.

Datasets. We collect responses from 50 LLMs on 20 strategic games. This leads to $N = 9,871$ valid response instances after filtering out failed, empty, or invalid responses.

Bottleneck Autoencoder Setup. Let i index response instances in dataset. Each response is encoded via `text-embedding-3-small` (OpenAI, 2024) to obtain an embedding vector $\mathbf{x}_i \in \mathbb{R}^{1536}$. We train a symmetric MLP autoencoder f_θ with widths:$1536-512-128-d-128-512-1536$. We use ReLU activations and Adam optimization, with a learning rate of 10^{-3}, batch size of 256, and 50 epochs. The autoencoder is trained to minimize the mean squared error (MSE) reconstruction loss:

$$\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N} \|\mathbf{x}_i - h_\psi(g_\phi(\mathbf{x}_i))\|^2, \tag{1}$$

where g_ϕ is the encoder and h_ψ the decoder.

Choice of Dimension d of Autoencoder. We use the explained variance ratio $\text{EVR}(d) = 1 - \text{MSE}_{\text{val}}(d)/\text{Var}_{\text{total}}$ to measure reconstruction quality at bottleneck dimension d, where $\text{MSE}_{\text{val}}(d)$ is the MSE reconstruction loss evaluated on the validation set, and $\text{Var}_{\text{total}}$ is the total validation variance. To find the value of d, we evaluate validation EVR across bottleneck widths and identify the first width at which the reconstruction curve has effectively plateaued. Let $\widetilde{\text{EVR}}(d)$ be the moving average of validation EVR across adjacent three bottleneck widths, and define

$$d^*(\epsilon_{\text{AE}}) = \min\{d : \widetilde{\text{EVR}}(d) \geq \max_{d'} \widetilde{\text{EVR}}(d') - \epsilon_{\text{AE}}\}. \tag{2}$$

We set $\epsilon_{\text{AE}} = 1\%$ by default. Thus, d^* is the smallest bottleneck width whose smoothed validation EVR is within one percentage point of the best observed smoothed EVR. To assess robustness of the inferred latent geometry, we evaluate a grid of 12 autoencoder configurations spanning 3 widths, 2 activations (ReLU, GELU), and 2 weight decay values (0, 10^{-5}), with 5 seeds each. Training details and convergence analysis are in Appendix C.

Behavioral Signatures. Let m index models and $q \in \mathcal{G}$ index games, and let $\mathbf{x}_{m,q} \in \mathbb{R}^{1536}$ denote the embedding of model m's response to game q. Each model m receives a behavioral signature

$$\bar{\mathbf{z}}_{m,q} = \frac{1}{|\mathcal{I}_{m,q}|}\sum_{i \in \mathcal{I}_{m,q}} g_\phi(\mathbf{x}_i), \qquad \mathbf{z}_m = \frac{1}{|\mathcal{G}_m|}\sum_{q \in \mathcal{G}_m} \bar{\mathbf{z}}_{m,q}, \tag{3}$$

where $\mathcal{I}_{m,q}$ is the set of responses by model m for game q and $\mathcal{G}_m$ denotes the set of evaluated game instances for which model m produced a valid response.

Dimensionality Metrics. We use the PCA-based dimension metric as our primary choice for quantifying the geometry of strategic responses. Given a set of response embeddings $\{\mathbf{x}_i\}_{i=1}^N$, we stack their transposes row-wise $X = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N]^\top \in \mathbb{R}^{N \times 1536}$. Thus, each row of X is one model response and each column is one embedding coordinate. We center this matrix by subtracting the column-wise mean, obtaining $X_c = X - \mathbf{1}\bar{\mathbf{x}}^\top$, where $\bar{\mathbf{x}} = \frac{1}{N}\sum_{i=1}^N \mathbf{x}_i$. The covariance matrix is then $\Sigma = X_c^\top X_c/(N-1)$. Its eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots$ quantify how much response variance lies along each principal direction in the embedding space. For a target explained-variance threshold τ, we define

$$d_\tau = \min\left\{k : \frac{\sum_{j=1}^k \lambda_j}{\sum_j \lambda_j} \geq \tau\right\} \tag{4}$$

to be the smallest number of principal components needed to explain at least a fraction τ of total variance. Unless otherwise stated, we report d_{90} as our PCA-based dimensionality metric. We evaluate d_{90} under three complementary data constructions:

- **Balanced PCA** computes d_{90} on balanced subsets of n LLMs, asking whether adding more LLMs expands the space of strategic responses or mainly adds samples inside an existing subspace. We sample n LLMs without replacement. We then balance the sampled response matrix so that each model contributes the same number of valid responses, compute d_{90} using Eqn. 4, and repeat this procedure 10 times. The reported value is the mean over repeats, with variability across repeats used to assess sampling stability.
- **Full pooled PCA** computes d_{90} on all individual response embeddings to measure total response-space dimensionality. This case measures the total linear dimensionality of all observed strategic responses, including both differences across games and differences across models. It is evaluated by applying Eqn. 4 once to the centered matrix of all valid response embeddings $\{\mathbf{x}_i\}_{i=1}^{N}$. This metric answers how many linear directions are needed to reconstruct the overall response cloud at 90% explained variance.
- **Model-means PCA** computes d_{90} after averaging responses within each model to isolate between-model variation. This case isolates persistent between-model variation from within-model, per-game variation. For each model, we first compute its mean response embedding $\bar{\mathbf{x}}_m = \frac{1}{|\mathcal{G}_m|} \sum_{q \in \mathcal{G}_m} \mathbf{x}_{m,q}$, where $\mathcal{G}_m$ is the set of games with valid responses for model m. We then apply Eqn. 4 to the 45 model means $\{\bar{\mathbf{x}}_m\}_{m=1}^{45}$. The resulting dimension estimates the number of independent directions along which models differ after averaging over task-specific response variation. This is distinct from Eqn. 3 which is used for the game-level alignment analysis.

3 Response-Space Expansion Exhibits Diminishing Returns

We evaluate how strategic diversity scales across LLMs by measuring d_{90} as a function of number of models n. We consider nine choices of n in $\{5, 10, 15, 20, 25, 30, 35, 40, 45, 50\}$ and compute d_{90} on balanced subsets of n LLMs. We run 10 trials for each value of n.

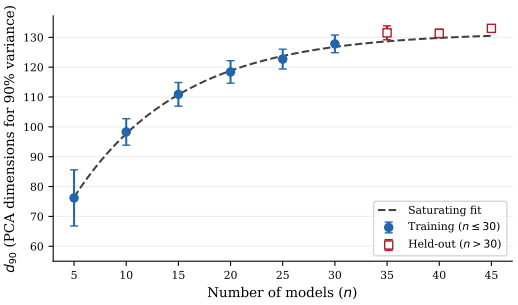

Figure 2: Sublinear scaling of ecosystem dimensionality. This figure shows balanced PCA d_{90} versus ecosystem size n.

Finding: Pooled response-space dimensionality grows sublinearly over the observed ecosystem. Figure 2 shows that d_{90} rises steeply at small n and flattens as n grows: from 76.2 ($n = 5$) to 133.0 ($n = 45$), with standard deviations decreasing from ± 9.4 to 0. Moreover, the marginal slope declines monotonically, decreasing from 3.47 dims/model between $n = 5$ and $n = 15$ to 0.35 dims/model between $n = 30$ and $n = 45$. Spearman rank correlation between n and marginal slope is $\rho = -0.95$ ($p = 0.0003$), confirming that d_{90} grows sublinearly rather than linearly with ecosystem size $n \leq 45$. These results show that additional models contribute progressively less expansion to the total observed response space. Because this analysis includes variation across games, instances, and repeated generations, we will investigate between-model behavioral dimensionality using model-level signatures in Section 4.

Finding: Balanced PCA dimensionality follows a saturating exponential scaling law. To further investigate the saturation pattern of balanced PCA, we fit candidate scaling laws using the subset $n \leq 30$. We find that the saturating exponential form $d(n) = M(1 - e^{-kn/M}) + c$ best captures the observed pattern with parameters $M = 89.8$, $k = 8.7$, $c = 41.8$, outperforming other possible patterns including linear, logarithmic, and power-law alternatives by Bayesian information criterion (Schwarz, 1978). We test the fitted model by extrapolating to $n = 45$ and comparing to the observed $d_{90}(45)$ in Figure 2. We find that the extrapolated value $d_{90}(45) = 130.5$ falls within -1.9% of the observed 133.0, with a bootstrap 95% confidence interval of $[122.0, 147.4]$, suggesting that the saturating form captures genuine structural regularity rather than overfitting noise in the training data. Extrapolating the fitted curve beyond $n = 45$ predicts little additional growth: the model approaches

an asymptote of $M + c \approx 132$, suggesting that further models likely populate the existing response space rather than introduce many new principal directions.

4 The Bounded Geometry Beneath Diminishing Returns

We have shown that strategic diversity saturates under balanced PCA. This saturation raises a structural question: *are additional models failing to add new strategic degrees of freedom, or is PCA merely flattening a richer nonlinear geometry?* To test this, we run a reconstruction experiment with bottleneck autoencoders. Our insight is that if strategic responses keep requiring new nonlinear degrees of freedom as n grows, reconstruction quality should continue improving at larger bottleneck widths d. If the response space is bounded, reconstruction should instead reach a stable plateau. That is, beyond some bottleneck size d, extra dimensions should recover little additional structure. We use this empirical plateau to estimate the size of the *nonlinear core*: the effective nonlinear degrees of freedom needed to reconstruct strategic responses. By Eqn. 2, the smallest tested bottleneck width $d^* = 18$ with $\epsilon_{AE} = 0.01$.

4.1 Experiment Setup to Estimate the Nonlinear Core

To estimate the nonlinear core, we train autoencoders for each n to reconstruct response embeddings $\mathbf{x}$. Following Eqn. (2), we estimate $d^*(\epsilon_{\text{AE}})$ as the smallest bottleneck dimension whose smoothed validation EVR is within ϵ_{AE} of the best observed smoothed EVR. Once EVR stops improving meaningfully, additional dimensions no longer recover new structure, indicating that the autoencoder has reached the effective nonlinear degrees of freedom in the data.

We sweep bottleneck dimensions $d \in \{8, 10, 12, 14, 16, 18, 20, 22, 24, 28, 32\}$ on both the $n = 30$ dataset (5,871 responses) and the $n = 45$ dataset (8,871 responses). By default, we run three random seeds (42, 123, 7). The default plateau tolerance is $\epsilon_{\text{AE}} = 0.01$.

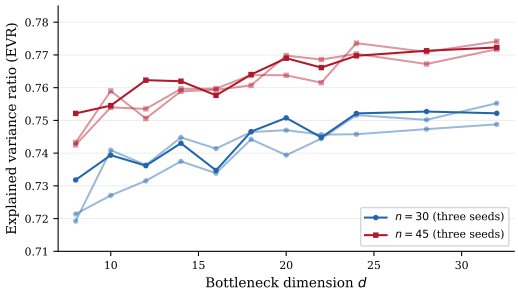

Figure 3: Nonlinear core invariance across ecosystem sizes. This figure shows EVR vs. bottleneck d for $n = 30$ (blue) and $n = 45$ (red) across 3 seeds each. Under the primary ReLU protocol, Eqn. 2 selects both plateau onset of $d^* = 18$.

4.2 Experimental Results and Interpretation

Finding: The estimated nonlinear core dimension is invariant as n varies. Figure 3 reports EVR as a function of bottleneck dimension. Under the ReLU activation, the median plateau occurs at $d^* = 18$ for both $n = 30$ and $n = 45$. Thus, adding 15 additional models does not increase the onset of the nonlinear plateau.

Table 1: PCA metrics at $n = 30$, $n = 45$, and the expanded $n = 50$ endpoint.

Metric	$n = 30$	$n = 45$	$n = 50$
Pooled PCA d_{90}	129	133	137
Balanced PCA d_{90}	127.8	133.0	136.2
Model-means PCA d_{90}	12	16	16

Finding: Added models fill in the bounded core rather than expanding it. Figure 3 shows that at every bottleneck width, the $n = 45$ autoencoder achieves higher EVR than the $n = 30$ autoencoder. This uniform EVR gain is consistent with denser coverage of the same bounded nonlinear structure. The PCA-based metrics in Table 1 show what the additional models contribute. As we will rule out the within-model response variation later, the increase in PCA-based metrics indicates genuinely new between-model orientations. Combining the discussion above, added models are not redundant, but their novelty appears as denser coverage and new linear orientations within a bounded nonlinear core, rather than as new nonlinear degrees of freedom.

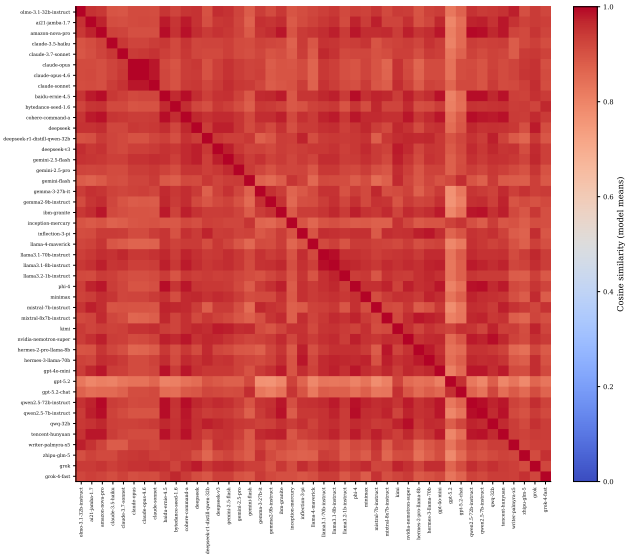

Figure 4: This figure shows inter-model cosine similarity across 45 models from 26 families. Models exhibit mean similarity 0.927. Stripping responses to final decisions (8% of text) preserves the pattern (Appendix H).

Finding: Game specific cores are bounded. To test whether the plateau estimated on the dataset from game suite masks a high-dimensional substructure in any particular game type, we repeat the experiments within each game category on the $n = 45$ dataset and report medians over 15 seeds. When $\epsilon_{\text{AE}} = 0.01$, we find bounded plateau dimensions associated with each game category: Signaling $d^* = 12$ (12.9 ± 5.3), Deception $d^* = 16$ (16.7 ± 3.8), Bayesian $d^* = 20$ (20.8 ± 3.9), and Coordination $d^* = 18$ (18.4 ± 4.0). No category's median requires more than 20 bottleneck dimensions to reach its reconstruction plateau, matching the scale of the bounded dimensional plateau estimated from the dataset of game suite. Thus, the bounded nonlinear core is not an averaging artifact caused by mixing low-dimensional categories with one high-dimensional category. Figure 4 visualizes overlap in shared response geometry. Across 45 models, the mean inter-model cosine similarity is 0.927.

Finding: Between-model structure is concentrated at the family level. When model families are added greedily according to their marginal contribution, PCA recomputed on the models from the first five selected families yields $d_{90} = 9$, compared with $d_{90} = 16$ for the complete 45-model ecosystem. Subsequent families therefore expand the between-model representation by only seven additional dimensions, with many producing no change in the cumulative d_{90}. Many subsequent families produce no increase in cumulative d_{90}, although the complete ecosystem still expands from 9 to 16 dimensions. Variance decomposition further shows that most overall variation occurs within models across games (93.4%), while only 6.6% occurs between models. However, this between-model component is the relevant source of model-level diversity, and 62.4% of it lies between families. Thus, for workflows seeking complementary model behaviors, selecting models from distinct families is more effective than simply increasing the number of models. PC-band profiles also vary substantially within families, so family remains an imperfect proxy for behavioral geometry (Appendix E').

Ablations Studies. We perform ablation studies on task domains, autoencoder configurations, and datasets in Appendix C, F, and L, respectively.

4.3 Models Share a Behavioral Coordinate System

The bounded nonlinear core shows that strategic behavior occupies a compact geometry, but it does not yet establish whether different models organize that geometry in compatible ways. We therefore test whether game-level behavioral signatures align across model families.

Setup and Metric. For each model m and game $q \in \mathcal{G}$, we compute the centroid of its response embeddings $\bar{\mathbf{z}}_{m,q} = \frac{1}{|\mathcal{I}_{m,q}|} \sum_{i \in \mathcal{I}_{m,q}} g_\phi(\mathbf{x}_i)$. Stacking these centroids across games gives a game-level behavioral signature matrix for each model. We compare signature matrices between models using orthogonal Procrustes alignment and test whether the aligned representations preserve game identity across model families.

Cross-family geometric alignment. The bounded nonlinear core does not by itself establish whether different models organize strategic games in compatible ways. A complementary game-level analysis finds strong cross-model alignment of behavioral signatures, including similar within-family

and across-family alignment. These results suggest that models occupy not only a bounded response space, but a similarly organized one. Because the alignment is evaluated on the same games used to estimate the transformation, we treat it as descriptive evidence and report the full analysis and limitations in Appendix D.

5 Aggregation Does Not Necessarily Expand Strategic Diversity

5.1 Experiment Setup

The preceding sections show that diversity saturates as models are added. Here we test three common hopes for overcoming that limit: (1) majority vote, (2) multi agent discussion, and (3) cross provider composition. In each case, we ask whether aggregation produces decisions outside the shared response geometry, rather than merely increasing agreement within it.

Aggregation patterns. We evaluate three aggregation patterns using strategic scores. The strategic score is one when the response's final action matches the game's target action, and zero otherwise.

- **Majority vote.** We evaluate GPT 5.2, Claude Opus 4.6, Gemini 2.5 Flash, Llama 3.1 70B, and DeepSeek V3. Each model answers the same game as many as 50 times. We use K to denote the number of responses combined into one decision. For example, $K = 1$ uses a single response, while $K = 50$ selects the action that appears most often among 50 responses. We evaluate $K \in \{1, 3, 5, 10, 20, 50\}$. For every value of K, we form 200 matched random samples from each response pool. Six games have a well defined target action, giving 30 combinations of models and games for the strategic score analysis.
- **Multi agent discussion.** We analyze three teams of three models: a team containing three instances of GPT 5.2, a team of frontier models from different providers, and a team mixing closed and open models (Appendix J). Each team discusses every game for three rounds, with two trials per game and 40 trials per team. We ask two questions. First, do the models become more similar to one another? Second, do their answers move beyond the behaviors already observed in the ecosystem? We measure this second effect by distance from the shared PCA subspace and distance from the nearest independent response to the same game. We also measure strategic score whenever the game has a well defined target.
- **Cross provider composition.** One model proposes a strategy, a model from another provider critiques it, and the first model then chooses a final action. We use GPT 5.6, Claude Opus 5, GLM 5.2, Qwen 3.8 Max, and DeepSeek V3. We form five provider pairs and test both role assignments on five games, with five trials for each assignment (Appendix K). This produces 250 transcripts, of which 241 have valid scores for both the proposal and final answer. We compare each final answer with its proposal to determine whether critique changes the action and whether the change improves the score.

5.2 Experiment Results

Finding: Majority voting helps when the stronger action is already common. With one response, the mean strategic score is 0.492. Combining 50 responses raises it to 0.600. The mean gain is 0.108, with a 95% confidence interval from 0.020 to 0.207 across the 30 combinations of models and games.

The gain does not mean that voting discovers a new strategy. Voting gives more weight to the action that a model already produces most often. This helps when that action is strong and reinforces an error when it is not. At $K = 50$, 15 combinations improve, nine receive a lower score, and six do not change. When results are averaged by game, two games improve and four show no clear change. Majority voting is therefore useful for recovering a model's dominant strategic tendency, but its value depends on the quality of that tendency.

Finding: Diverse discussion creates novelty, but novelty does not guarantee a better decision. The team containing three instances of GPT 5.2 becomes more internally similar. Its mean pairwise similarity rises from 0.856 in the first round to 0.916 in the third round. The increase is 0.064, with a 95% confidence interval from 0.020 to 0.109 and $p = 0.010$. At the same time, this team's distance

from the shared subspace and from the nearest independent response does not clearly change. A homogeneous discussion therefore produces stronger agreement without producing a new direction.

The two diverse teams show the complementary pattern. Their members do not become clearly more similar, but their answers become more novel. For the frontier team, distance from the shared subspace and from the nearest independent response each increase by 0.050. For the team mixing closed and open models, the increases are 0.018 and 0.035. All four changes have $p = 0.005$. Strategic score, however, shows no clear improvement for any team. Diversity among participants can move discussion beyond familiar responses, but an unfamiliar response is not necessarily a strategically useful one.

Finding: Cross provider critique changes decisions more often than it improves them. Critique has a visible effect on behavior. The final action differs from the proposal in 97 of 241 transcripts, or 40.2%. When the critic explicitly states a verdict, it disagrees with the proposal in 74 of 153 cases. Among these disagreements, the final action changes in 52 cases, or 70.3%.

Most of these revisions do not improve the strategic outcome. The score is unchanged in 209 of 241 transcripts, or 86.7%, and improves in 11 transcripts, or 4.6%. The mean score is 0.207 before critique and 0.166 afterward. The estimated decrease is 0.041, but the 95% confidence interval from -0.087 to 0.004 includes zero, with $p = 0.117$. Changing which model proposes the answer also has little effect: reciprocal role assignments produce the same final score in 97 of 117 matched cases. Cross provider critique can redirect a decision, but it does not reliably identify a better direction.

Boundary check beyond strategic games. The bounded geometry is not unique to our game suite: on open-ended prompts, the same 45-model ecosystem yields an identical nonlinear plateau ($d^* = 18$) and nearly unchanged between-model dimensionality (model-means PCA $d_{90} = 17$ versus 16 for strategic games). This transfer supports interpreting the aggregation results as consequences of a shared ecosystem geometry; Appendix L provides the cross-domain design, response-level contrast, and full results.

6 Conclusion

We evaluated 50 large language models from 26 model families on a common suite of strategic-game conditions and examined how the geometry of their responses changed as the evaluated ecosystem grew. The total response space expanded with sharply diminishing marginal returns, while PCA over model-level behavioral signatures required 16 components to explain 90% of between-model variance at $n = 45$. Adequately expressive autoencoders exhibited architecture-dependent plateaus between 18 and 24 bottleneck dimensions, and model signatures showed substantial similarity and cross-family alignment. These findings imply that nominal ecosystem growth produces less behavioral expansion than model and provider counts alone suggest. For collective AI systems, the central design objective should therefore be measured complementarity. Additional models are valuable when they contribute behavioral directions absent from the existing system, not merely when they carry a different name or provider label.

AI Use Statement

In this work, we used generative AI tools for generating synthetic data sets, implementing methods, and cleaning and reformatting dataset. We have not used generative AI tools for proposing or refining hypotheses, designing or providing feedback on research methodology or experiment, interpreting results and helping develop theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, supporting qualitative and thematic data analysis are not applicable to this work. Additionally, we used generative AI tools for creating illustrative images. We have reviewed all AI-assisted work by reviewing and testing LLM-generated code for correctness. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

Ethics Statement

This work studies the strategic behavior of large language models using synthetic, self-contained game scenarios. No human participants, personally identifiable information, private communications, or sensitive personal data were involved. Some evaluated scenarios involve deception, signaling, negotiation, and strategic manipulation. Although these controlled tasks may reveal transferable patterns in deceptive reasoning, we present the results to characterize correlated model failures and improve the reliability of ensemble and multi-agent systems. The proposed geometric measurements should not be treated as complete safety evaluations or used as the sole basis for high-stakes deployment decisions: their validity depends on the evaluated models, prompts, games, embedding models, and analysis procedures. In particular, low-dimensional similarity does not establish that models are interchangeable or that an unevaluated behavior is safe. Researchers releasing or reusing the response data should comply with the applicable model-provider terms and licenses.

Reproducibility Statement

We will provide the information and artifacts needed to reproduce the study. Section 2 specifies the dataset construction, embedding procedure, autoencoder architecture, optimization settings, dimensionality metrics, and plateau-selection rule. Appendix A documents the strategic games and prompt design; Appendix B lists the evaluated models and query configuration; and Appendix C reports the training protocol, hardware requirements, and random-seed settings. Additional robustness procedures are described in Appendices F, H, and I. Because hosted models and APIs may change over time, exact replication additionally depends on access to the model versions identified in Appendix B.

References

Elif Akata, Lion Schulz, Julian Coda-Forno, Seong Joon Oh, Matthias Bethge, and Eric Schulz. Playing repeated games with large language models. *Nature Human Behaviour*, 9(7):1380–1390, 2025.

Philip Brookins and Jason M DeBacker. Playing games with GPT: What can we learn about a large language model from canonical strategic games? *SSRN*, 2023.

John P Cunningham and Byron M Yu. Dimensionality reduction for large-scale neural recordings. *Nature Neuroscience*, 17(11):1500–1509, 2014.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, et al. Toy models of superposition. *arXiv preprint arXiv:2209.10652*, 2022.

Caoyun Fan, Jindou Chen, Yaohui Jin, and Hao He. Can large language models serve as rational players in game theory? A systematic analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024.

Kanishk Gandhi, Dorsa Sadigh, and Noah D Goodman. Strategic reasoning with language models. *arXiv preprint arXiv:2305.19165*, 2023.

Fulin Guo. GPT in game theory experiments. *arXiv preprint arXiv:2305.05516*, 2023.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. In *International Conference on Learning Representations*, 2023.

Mehrdad Jazayeri and Srdjan Ostojic. Interpreting neural computations by examining intrinsic and embedding dimensionality of neural activity. *Current Opinion in Neurobiology*, 70:113–120, 2021.

Dongfu Jiang, Xiang Ren, and Bill Yuchen Lin. Llm-blender: Ensembling large language models with pairwise ranking and generative fusion. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14165–14178, 2023.

Liwei Jiang, Yuanjun Chai, Margaret Li, Mickel Liu, Raymond Fok, Nouha Dziri, Yulia Tsvetkov, Maarten Sap, Alon Albalak, and Yejin Choi. Artificial hivemind: The open-ended homogeneity of language models (and beyond). In *Advances in Neural Information Processing Systems*, 2025.

Elizaveta Levina and Peter J Bickel. Maximum likelihood estimation of intrinsic dimension. *Advances in Neural Information Processing Systems*, 17, 2004.

Isaac Ong, Amjad Almahairi, Vincent Wu, Wei-Lin Chiang, Tianhao Wu, Joseph E Gonzalez, M Waleed Kadous, and Ion Stoica. Routellm: Learning to route llms with preference data. *International Conference on Learning Representations (ICLR)*, 2025.

OpenAI. New embedding models and API updates. `https://openai.com/index/new-embedding-models-and-api-updates/`, January 2024.

Gideon Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, pp. 461–464, 1978.

Hongjin Su, Weijia Shi, Jungo Kasai, Yizhong Wang, Yushi Hu, Mari Ostendorf, Wen-tau Yih, Noah A Smith, Luke Zettlemoyer, and Tao Yu. One embedder, any task: Instruction-finetuned text embeddings. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 1102–1121, 2023.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L Turner, Callum McDougall, Monte MacDiarmid, C Daniel Freeman, Theodore R Sumers, Edward Rees, Joshua Batson, Adam Jermyn, Shan Carter, Chris Olah, and Tom Henighan. Scaling monosemanticity: Extracting interpretable features from Claude 3 Sonnet. *arXiv preprint arXiv:2605.29358*, 2026.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*, 2022.

Wikimedia Foundation. Wikimedia downloads: Database backup dumps. `https://dumps.wikimedia.org/`, 2026.

Mitchell Wortsman, Gabriel Ilharco, Samir Yitzhak Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carber, Simon Kornblith, and Ludwig Schmidt. Model soups: Averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International Conference on Machine Learning*, 2022.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, et al. Autogen: Enabling next-gen llm applications via multi-agent conversations. In *First conference on language modeling*, 2024.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin Raffel, and Mohit Bansal. TIES-merging: Resolving interference when merging models. In *Advances in Neural Information Processing Systems*, 2023.

Le Yu, Bowen Yu, Haiyang Yu, Fei Huang, and Yongbin Li. Language models are Super Mario: Absorbing abilities from homologous models as a free lunch. In *Forty-first International Conference on Machine Learning*, 2024.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J Zico Kolter, and Dan Hendrycks. Representation engineering: A top-down approach to AI transparency. In *arXiv preprint arXiv:2310.01405*, 2023.

APPENDIX

A GAME DESCRIPTIONS

The 20 strategic games are designed to probe four facets of strategic reasoning. Each game is presented as a natural-language scenario with explicit payoff structures and unambiguous action spaces. We deliberately span a wide range of game-theoretic concepts (incomplete information, repeated interaction, signaling, coordination under uncertainty, Bayesian updating, and strategic deception) to ensure that the resulting embeddings capture broad strategic competence.

Selection rationale. The game suite is constructed to maximize coverage of strategic motifs. Each category probes a distinct facet of strategic reasoning (communication, alignment, updating, misrepresentation), and within each category the five games are designed to vary along three axes: action space (binary, discrete, continuous), information structure (complete, incomplete, private), and equilibrium concept (pure Nash, mixed, Bayesian, subgame-perfect). Table 2 summarizes this coverage. We begin from game templates, where each template specifies the strategic structure: players, action spaces, information structure, and payoff logic. Seed variants are generated by re-instantiating selected templates with new sampled factors, such as payoff values, private types or signals, role assignments, or surface scenario details, depending on the game, while preserving the underlying strategic structure. The resulting 20-game suite combines distinct templates with four such seed variants, marked with * in Table 2.

Table 2: **Game coverage across strategic motifs.** The 20 games span 4 categories, 3 action space types, 3 information structures, and 5 equilibrium concepts, ensuring broad coverage of game-theoretic reasoning. Games marked with * are independently generated seed variants of an earlier game template.

Game	Action Space	Information	Equilibrium
Signaling			
Job Market Signaling	Discrete (0–3)	Incomplete	Separating/Pooling
Credential Signaling	Discrete (4)	Incomplete	Separating/Pooling
Warranty Signaling	Discrete (0–3)	Incomplete	Separating/Pooling
Job Market Signaling*	Discrete (0–3)	Incomplete	Separating/Pooling
Credential Signaling*	Discrete (4)	Incomplete	Separating/Pooling
Coordination			
Technology Adoption	Discrete (4)	Complete	Pure Nash (multiple)
Battle of the Sexes	Binary	Complete	Mixed Nash
Stag Hunt	Binary	Complete	Pure Nash (risk)
Weak Link Game	Discrete (1–7)	Complete	Pure Nash
Battle of the Sexes*	Binary	Complete	Mixed Nash
Deception			
Bluffing Poker	Binary	Private card	Mixed Nash
Mafia Game	Message + Vote	Private role	Bayesian
Negotiation	Continuous (offers)	Private value	Subgame-perfect
Spy Game	Vote + Action	Private role	Bayesian
Two-Card Monte	Binary	Private card	Mixed Nash
Bayesian			
First-Price Auction	Continuous (bid)	Private value	Bayes–Nash
Common Prior Game	Discrete (3)	Private signal	Bayesian
Global Game	Binary	Private signal	Global game eqm
Dating Market	Discrete (3)	Incomplete	Signaling
First-Price Auction*	Continuous (bid)	Private value	Bayes–Nash

The descriptions below cover the full 20-game suite. Split-specific use of these games is discussed separately in the training and cross-validation protocol (Appendix C); the game descriptions themselves are not intended to imply a substantive hierarchy between training and held-out games. Within each description, printed braces {...} mark fields whose values can be varied across independently generated seed instances.

Signaling: Job Market Signaling.

- **Initial template**: A worker has private productivity—{high (worth \$100)} or {low (worth \$60)}, each with probability {0.5}—and education costs differ by type: {\$20/year} for high-productivity workers and {\$40/year} for low-productivity workers. The employer observes education and offers a wage; the worker's payoff is wage minus education cost, and the employer's payoff is true productivity minus wage.
- **Type**: Costly signaling.
- **Players**: One worker and one employer.
- **Action space**: The worker chooses education from $\{0, 1, 2, 3\}$ years; the employer offers a wage after observing education.
- **Information structure**: The worker privately observes productivity; the employer observes only education.
- **Equilibrium concept**: Separating or pooling equilibrium, with the probe focused on whether education credibly separates high- from low-productivity workers.

Signaling: Credential Signaling.

- **Initial template**: A worker has private ability—{high (worth \$120)} or {low (worth \$60)}, each with probability {0.5}. Credential costs are {[0, 15, 35, 60]} for high-ability and {[0, 30, 70, 120]} for low-ability workers, with the cost gap widening at higher tiers.
- **Type**: Multi-level costly signaling.
- **Players**: One worker and one employer.
- **Action space**: The worker chooses a credential from {BA, MA, PhD, Post-doc}; the employer observes the credential and offers a salary.
- **Information structure**: The worker privately observes ability; the employer observes only the credential.
- **Equilibrium concept**: Separating or pooling equilibrium, with the probe focused on whether advanced credentials are credible because they are differentially costly.

Signaling: Warranty Signaling.

- **Initial template**: A seller has a product of private quality—{high (worth \$100)} or {low (worth \$40)}, each with probability {0.5}. Warranty costs are {\$5/year} for high-quality products and {\$20/year} for low-quality products.
- **Type**: Costly quality signaling.
- **Players**: One seller and one buyer.
- **Action space**: The seller offers a warranty of $\{0, 1, 2, 3\}$ years; the buyer observes warranty length and offers a price.
- **Information structure**: The seller privately observes product quality; the buyer observes only warranty length.
- **Equilibrium concept**: Separating or pooling equilibrium, with the probe focused on whether warranty length credibly signals quality.

Signaling: Job Market Signaling.

- **Initial template**: Same strategic template as Job Market Signaling, re-instantiated with independently generated seed factors {productivity values, type prior, education costs, role labels, surface scenario details} while preserving the differential education-cost structure.
- **Type**: Costly signaling seed variant.
- **Players**: One worker and one employer.
- **Action space**: The worker chooses education from $\{0, 1, 2, 3\}$ years; the employer offers a wage after observing education.

- **Information structure**: The worker privately observes productivity; the employer observes only education.
- **Equilibrium concept**: Separating or pooling equilibrium.

Signaling: Credential Signaling.

- **Initial template**: Same strategic template as Credential Signaling, re-instantiated with independently generated seed factors {ability values, type prior, credential labels, credential costs, surface scenario details} while preserving the multi-level credential-cost structure.
- **Type**: Multi-level costly signaling seed variant.
- **Players**: One worker and one employer.
- **Action space**: The worker chooses a credential from {BA, MA, PhD, Post-doc}; the employer offers a salary after observing the credential.
- **Information structure**: The worker privately observes ability; the employer observes only the credential.
- **Equilibrium concept**: Separating or pooling equilibrium.

Coordination: Technology Adoption.

- **Initial template**: Three players choose among symmetric technology standards {A, B, C, D}. All options yield a base payoff of {\$10} plus a {\$15} network effect for each other player choosing the same option.
- **Type**: Complete-information coordination with multiple symmetric equilibria.
- **Players**: Three technology adopters.
- **Action space**: Each player independently chooses a standard from {A, B, C, D}.
- **Information structure**: Complete information about payoffs and available standards.
- **Equilibrium concept**: Pure-strategy Nash equilibrium; all players coordinating on the same standard are equilibria, and the probe is whether a Schelling focal point emerges.

Coordination: Battle of the Sexes.

- **Initial template**: Two players prefer coordinating over miscoordinating but disagree about the best coordinated outcome. Coordinated activities are {Opera} and {Football}; payoffs are {(Opera, Opera) = $(3, 2)$}, {(Football, Football) = $(2, 3)$}, and mismatches yield {$(0, 0)$}.
- **Type**: Coordination with conflicting preferences.
- **Players**: Two players.
- **Action space**: Each player independently chooses Opera or Football.
- **Information structure**: Complete information about both players' preferences and payoffs.
- **Equilibrium concept**: Two pure-strategy Nash equilibria and one mixed Nash equilibrium, with the probe focused on navigating divergent preferences.

Coordination: Stag Hunt.

- **Initial template**: Two hunters choose between a payoff-dominant cooperative action {Stag} and a safer individual action {Hare}. Payoffs are {(Stag, Stag) = $(4, 4)$}, {(Hare, Hare) = $(3, 3)$}, {(Stag, Hare) = $(0, 3)$}, and {(Hare, Stag) = $(3, 0)$}.
- **Type**: Trust-based coordination.
- **Players**: Two hunters.
- **Action space**: Each player independently chooses Stag or Hare.
- **Information structure**: Complete information about payoffs and actions.
- **Equilibrium concept**: Pure-strategy Nash equilibrium, with a payoff-dominant equilibrium at (Stag, Stag) and a risk-dominant equilibrium at (Hare, Hare).

Deception: Simplified Poker.

- **Initial template**: A Kuhn-poker variant uses three cards $\{\text{Low} = 1, \text{Medium} = 2, \text{High} = 3\}$, dealt without replacement to two players, with a {\$1 ante} each.
- **Type**: Strategic deception and bluffing.
- **Players**: Two poker players.
- **Action space**: Player 1 chooses Bet (\$2) or Check; Player 2 responds with Call/Fold or Check as appropriate.
- **Information structure**: Each player privately observes their own card and reasons under uncertainty about the opponent's card.
- **Equilibrium concept**: Mixed Nash equilibrium, with the probe focused on bluffing weak hands and value-betting strong hands.

Deception: Mafia.

- **Initial template**: Three players are assigned roles: {two Villagers} and {one Mafia member}. Players make public statements and then vote to eliminate one player; Villagers earn $\{+5\}$ if Mafia is eliminated $\{(-5$ otherwise)$\}$, while Mafia earns $\{+10\}$ if not eliminated $\{(-10$ if eliminated)$\}$.
- **Type**: Hidden-role deception.
- **Players**: Three players: two Villagers and one Mafia member.
- **Action space**: Each player chooses a public statement and then votes to eliminate one player.
- **Information structure**: Each player observes their own role; role information is otherwise private.
- **Equilibrium concept**: Bayesian equilibrium, with the probe focused on deceptive role claims and inference from statements.

Deception: Buyer–Seller Negotiation.

- **Initial template**: A seller has a private cost from {[\$20, \$50]} and a buyer has a private value from {[\$40, \$80]}, with gains from trade guaranteed.
- **Type**: Bargaining with private values.
- **Players**: One buyer and one seller.
- **Action space**: Over 3 rounds of alternating offers, each party can state a price, accept, or walk away.
- **Information structure**: The seller privately observes cost and the buyer privately observes value.
- **Equilibrium concept**: Subgame-perfect equilibrium under private information, with the probe focused on reservation-price misrepresentation, anchoring, and credible walk-away threats.

Bayesian: First-Price Sealed-Bid Auction.

- **Initial template**: Two bidders each draw a private valuation uniformly from {[\$20, \$100]}. The highest bidder wins and pays their own bid.
- **Type**: Private-value auction.
- **Players**: Two bidders.
- **Action space**: Each bidder simultaneously submits a continuous sealed bid.
- **Information structure**: Each bidder privately observes their own valuation and knows the valuation distribution.
- **Equilibrium concept**: Bayes–Nash equilibrium; the strategic probe is bid shading, with equilibrium bidding function $b(v) = (v + 20)/2$.

Bayesian: Common Prior Game.

- **Initial template**: The true state is {Low, Medium, or High}, each equally likely. Each player receives a private signal matching the true state with {70%} probability; payoffs are $\{+10\}$ each if both correct, $\{-5\}$ each if both wrong, and $\{0\}$ each if split.
- **Type**: Bayesian updating with coordination incentives.
- **Players**: Two players.
- **Action space**: Each player independently guesses Low, Medium, or High.
- **Information structure**: Players share a common prior but receive conditionally informative private signals.
- **Equilibrium concept**: Bayesian equilibrium, with the probe focused on combining noisy private evidence with the incentive to coordinate.

Bayesian: Global Game.

- **Initial template**: An unknown fundamental θ is drawn uniformly from {[30, 70]}, and each player observes a private noisy signal {θ plus Gaussian noise, $\sigma = 10$}. An attack succeeds if at least {50%} attack AND $\{\theta > 50\}$; payoffs are success $\{+20\}$, failure $\{-10\}$, and abstaining $\{0\}$.
- **Type**: Global game with strategic uncertainty.
- **Players**: Two players.
- **Action space**: Each player independently chooses Attack or Not Attack.
- **Information structure**: Each player observes a private noisy signal about the underlying fundamental.
- **Equilibrium concept**: Global-game threshold equilibrium, with the probe focused on threshold reasoning under uncertainty.

Coordination: Weak Link Game.

- **Initial template**: Two workers choose effort, and payoffs depend on the minimum effort: payoff $= \{2\} \times \min(\text{both efforts}) - \{\text{own effort cost}\}$.
- **Type**: Coordination with weakest-link production.
- **Players**: Two workers.
- **Action space**: Each worker independently chooses effort from $\{1, \ldots, 7\}$.
- **Information structure**: Complete information about payoffs and available effort levels.
- **Equilibrium concept**: Pure-strategy Nash equilibrium; all (k, k) for $k \in \{1, \ldots, 7\}$ are equilibria, with $(7, 7)$ payoff-dominant and $(1, 1)$ risk-dominant.

Deception: Spy Game.

- **Initial template**: Three players are assigned hidden roles: {two Resistance} and {one Spy}. A leader proposes a {two-player mission team}, players vote to approve or reject, and the mission outcome depends on whether the Spy sabotages.
- **Type**: Hidden-role deception and team selection.
- **Players**: Three players: two Resistance members and one Spy.
- **Action space**: The leader proposes a mission team; players vote approve/reject; Resistance must choose Success on the mission, while the Spy chooses Success or Sabotage.
- **Information structure**: Each player observes their own role; role identities are otherwise private.
- **Equilibrium concept**: Bayesian equilibrium, with the probe focused on Spy infiltration and Resistance inference.

Bayesian: Dating Market.

- **Initial template**: Each agent has a private compatibility type—{High (100)}, {Medium (60)}, or {Low (20)}—and sends a costly signal. If both match, payoff is $\min(\text{both values}) -$ signal cost$-$ {\$5}.
- **Type**: Two-sided costly signaling.
- **Players**: Two dating-market agents.
- **Action space**: Each agent chooses a signal: dinner (\$30), coffee (\$10), or text (\$1), and then decides whether to match.
- **Information structure**: Each agent privately observes their own compatibility type and observes the other's signal.
- **Equilibrium concept**: Signaling equilibrium, with the probe focused on whether signal costs separate high- from lower-compatibility types.

Coordination: Battle of the Sexes.

- **Initial template**: Same strategic template as Battle of the Sexes, re-instantiated with independently generated seed factors {activity labels, payoff values, player preference ordering, surface scenario details} while preserving asymmetric preferences over coordinated outcomes.
- **Type**: Coordination with conflicting preferences seed variant.
- **Players**: Two players.
- **Action space**: Each player independently chooses one of two coordinated activities.
- **Information structure**: Complete information about both players' preferences and payoffs.
- **Equilibrium concept**: Two pure-strategy Nash equilibria and one mixed Nash equilibrium.

Deception: Two-Card Monte.

- **Initial template**: A dealer holds {an Ace and a King} and reveals the Ace's {initial position} to the player. The player bets {\$10} and guesses the Ace's final position after the dealer acts.
- **Type**: Strategic deception with hidden action.
- **Players**: One dealer and one player.
- **Action space**: The dealer chooses Honest (keep position) or Swap (secretly move Ace); the player guesses the Ace's final position.
- **Information structure**: The dealer observes whether the Ace was moved; the player observes only the initial position and must infer possible manipulation.
- **Equilibrium concept**: Mixed Nash equilibrium, with the probe focused on dealer randomization and player inference.

Bayesian: First-Price Auction.

- **Initial template**: Same strategic template as First-Price Auction, re-instantiated with independently generated seed factors {valuation distribution, payoff values, item/scenario labels, bidder roles} while preserving the private-value first-price auction structure.
- **Type**: Private-value auction seed variant.
- **Players**: Two bidders.
- **Action space**: Each bidder simultaneously submits a continuous sealed bid.
- **Information structure**: Each bidder privately observes their own valuation and knows the valuation distribution.
- **Equilibrium concept**: Bayes–Nash equilibrium with bid shading.

A.1 Prompt Design and Data Collection

Each game prompt follows a consistent template:

1. a narrative setting the scene,
2. explicit payoff structures (numerical where possible),
3. a clear action space,
4. the specific role assigned to the LLM, and
5. an instruction to explain reasoning before choosing an action.

Prompts are designed to be self-contained, requiring no external knowledge beyond basic arithmetic.

The study involves two overlapping datasets. The **full ecosystem** comprises 50 models from 26 families, each responding to all 20 games, yielding 9,871 valid response instances after filtering out empty or invalid responses.

B Model Details

B.1 Model Inventory

Table 3 lists all 50 models used for evaluations, grouped by family. Models marked with † form the 13-model interpolation subset used for composition and task-vector analyses (Section 4.3).

B.2 Query Configuration

All models are queried with the following shared configuration. We use temperature 0.7 by default. The maximum output tokens are 12,000 (16,000 for reasoning models). All models share the system prompt "You are playing a strategic game. Analyze the situation using game-theoretic reasoning. Consider the payoff structure, other players' incentives, and any relevant equilibrium concepts. Then clearly state your action or decision." We do not provide few-shot examples.

C Training Details

C.1 Autoencoder Architecture

The autoencoder used for the composition and task-vector analysis has the following architecture:

Component	Layer	Output dim
Encoder g_ϕ	Linear + ReLU	256
	Dropout (0.1)	256
	Linear + ReLU	128
	Linear	d
Decoder h_ψ	Linear + ReLU	128
	Dropout (0.1)	128
	Linear + ReLU	256
	Linear	1536

The bottleneck dimension d varies across experiments: $d \in \{4, 8, 12, 16, 24, 32\}$. The architecture is symmetric to avoid privileging encoding over decoding. We use ReLU activations throughout by default unless otherwise specified. Dropout rate of 0.1 is applied after the first hidden layer in both encoder and decoder to prevent overfitting at low dimension d.

Scaling study (45-model ecosystem). The autoencoder used for the scaling, threshold, and robustness analysis uses a wider architecture: $1536 \to 512 \to 128 \to d \to 128 \to 512 \to 1536$ (ReLU activations, no dropout, batch size 256, 50 epochs, 80/20 train/validation split). Wider hidden layers (512 vs. 256) accommodate the larger dataset (8,871 vs. 2,660 instances). Architecture robustness was assessed via the 12-configuration grid (Appendix F).

Table 3: Full 50-model ecosystem inventory. All models accessed via APIs across 26 independent families. †: 13-model interpolation subset.

Model	Family	Subset
Claude Opus 4.6†	Anthropic	†
Claude Opus 5†	Anthropic	†
Claude Opus†	Anthropic	†
Claude Sonnet†	Anthropic	†
Claude 3.5 Haiku	Anthropic	
Claude 3.7 Sonnet	Anthropic	
DeepSeek-V3†	DeepSeek	†
DeepSeek†	DeepSeek	†
DeepSeek-R1-Distill-32B	DeepSeek	
Gemini 2.5 Flash†	Google	†
Gemini 2.5 Pro†	Google	†
Gemini Flash†	Google	†
Gemma 2 9B	Gemma	
Gemma 3 27B	Gemma	
GPT-5.5†	OpenAI	†
GPT-5.6†	OpenAI	†
GPT-5.2†	OpenAI	†
GPT-5.2-Chat†	OpenAI	†
GPT-4o-mini	OpenAI	
Grok†	xAI	†
Grok 4 Fast	xAI	
Kimi†	Moonshot	†
MiniMax†	MiniMax	†
Llama 3.1 8B	Meta/Llama	
Llama 3.1 70B	Meta/Llama	
Llama 3.2 1B	Meta/Llama	
Llama 4 Maverick	Meta/Llama	
Mistral 7B	Mistral	
Mixtral 8x7B	Mistral	
Qwen 2.5 7B	Qwen	
Qwen 2.5 72B	Qwen	
QwQ 32B	Qwen	
Qwen 3.8 MAX	Qwen	
Hermes 2 Pro 8B	Nous/Hermes	
Hermes 3 70B	Nous/Hermes	
Phi-4	Microsoft	
OLMo 3.1 32B	AI2	
Command A	Cohere	
Jamba 1.7	AI21	
Nova Pro	Amazon	
ERNIE 4.5	Baidu	
Nemotron Super	NVIDIA	
Palmyra X5	Writer	
Inflection 3 Pi	Inflection	
GLM-5	Zhipu	
GLM-5.2	Zhipu	
Seed 1.6	ByteDance	
Hunyuan	Tencent	
Granite 4.0	IBM	
Mercury	Inception	

Table 4: **Generic-text control compared with LLM strategic responses.** Both datasets use the same response encoder, autoencoder architecture, bottleneck widths, and plateau-selection rule. Wikipedia paragraphs exhibit substantially broader and less aligned geometry than the LLM strategic responses.

Measure	Wikipedia paragraphs	LLM strategic responses
Autoencoder plateau d^*, 1% tolerance	22	18
Autoencoder plateau d^*, 0.5% tolerance	24	22
Peak validation EVR	0.41	0.76
Mean cosine similarity among group means	-0.02	0.93
PCA d_{90} of group means	31	16
Full pooled PCA d_{90}	416	133
Number of responses	9,000	8,871
Number of random seeds	15	5

For Wikipedia, groups correspond to the 45 topical categories. For the strategic-response dataset, groups correspond to the 45 evaluated LLMs. The 1% tolerance is the primary plateau-selection setting used in the main experiments.

C.2 Training Protocol

The following protocol applies to the 13-model interpolation study; the 45-model scaling study uses the parameters specified in Section 2.

- **Optimizer**: Adam ($\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$)
- **Learning rate**: 10^{-3}
- **Batch size**: 64
- **Epochs**: 200 for all configurations
- **Weight initialization**: Kaiming uniform
- **Loss function**: MSE reconstruction loss (Eqn. (1))
- **Regularization**: Dropout only without weight decay or bottleneck penalty
- **Random seeds**: 5 seeds per configuration; we report mean metrics.

C.3 Text Embedding

We use `text-embedding-3-small` model (OpenAI, 2024) to convert LLM responses into embedding vectors. Each LLM response is embedded as a single vector $\mathbf{x}_i \in \mathbb{R}^{1536}$. The mean embedding norm across all instances is 0.85 ± 0.12.

C.4 Wikipedia Null Baseline

As a text-encoder control, we embed 9,000 Wikipedia paragraphs sampled from Wikimedia database dumps (Wikimedia Foundation, 2026) (200 from each of 45 topical categories, filtered to 100–4,000 characters to approximate the distribution of LLM game responses) with the same `text-embedding-3-small` model and train the identical autoencoder architecture at the same bottleneck grid, with the same seeds and plateau rule.

Across the reported measures, Wikipedia paragraphs exhibit a different geometry from the LLM strategic responses. Their reconstruction curve plateaus at a larger bottleneck width, their full pooled PCA dimensionality is substantially higher, and their topical-category means are nearly orthogonal. These differences show that the bounded and highly aligned geometry observed for LLM responses is not a universal consequence of the response encoder or autoencoder procedure.

This generic-text control does not establish whether the bounded geometry is specific to strategic reasoning, because the Wikipedia paragraphs are not generated by the evaluated LLM ecosystem. Appendix L addresses that question by applying the same analysis to open-ended responses from the same 45 models.

D Cross-Family Alignment of Behavioral Geometry

The main text establishes that strategic responses occupy a bounded nonlinear geometry. Here, we examine whether models from different families organize the evaluated games in compatible ways within that geometry.

D.1 Game-Level Behavioral Signatures

For model m and game $q \in \mathcal{G}$, let $\mathcal{I}_{m,q}$ denote the set of valid responses produced by model m for game q. We define the game-level latent centroid

$$\bar{\mathbf{z}}_{m,q} = \frac{1}{|\mathcal{I}_{m,q}|} \sum_{i \in \mathcal{I}_{m,q}} g_\phi(\mathbf{x}_i), \tag{5}$$

where g_ϕ is the encoder of the primary 18-dimensional ReLU autoencoder.

Stacking the centroids for all games gives a game-level behavioral signature matrix

$$Z_m = \begin{bmatrix} \bar{\mathbf{z}}_{m,1}^\top \\ \vdots \\ \bar{\mathbf{z}}_{m,|\mathcal{G}|}^\top \end{bmatrix} \in \mathbb{R}^{|\mathcal{G}| \times 18}. \tag{6}$$

Before alignment, each matrix is centered by subtracting its mean centroid across games.

D.2 Orthogonal Procrustes Alignment

For a pair of models m and m', we estimate the orthogonal transformation

$$R^*_{m,m'} = \arg\min_{R^\top R = I} \|Z_m - Z_{m'} R\|_F^2 . \tag{7}$$

We summarize alignment quality using

$$R^2_{\text{align}} = 1 - \frac{\left\|Z_m - Z_{m'} R^*_{m,m'}\right\|_F^2}{\|Z_m\|_F^2}. \tag{8}$$

Higher values indicate that the relative organization of games for one model can be closely matched to that of another model through an orthogonal transformation.

Aligning each model's game-level signature matrix to the ecosystem-average reference yields a mean Procrustes alignment of $R^2_{\text{align}} = 0.944 \pm 0.004$ across three autoencoder seeds. In separate direct model-to-model comparisons, mean alignment is 0.917 for the 32 within-family pairs and 0.900 for the 958 across-family pairs, corresponding to an overall pairwise mean of approximately 0.901. Thus, across-family alignment is only 0.017 lower than within-family alignment.

As a complementary descriptive test, we align the signature matrix of one model to that of another and retrieve the nearest game centroid. The corresponding game is retrieved with 80.9% accuracy among 20 candidates, compared with a 5% chance baseline.

D.3 Category-Level Shift Directions

We next examine whether broad category contrasts have similar directions across models. For model m and category c, define

$$\mu_{m,c} = \frac{1}{|\mathcal{G}_c|} \sum_{q \in \mathcal{G}_c} \bar{\mathbf{z}}_{m,q}, \tag{9}$$

where $\mathcal{G}_c$ is the set of games belonging to category c. For two categories c_+ and c_-, the model-specific shift is

$$v_m^{c_+,c_-} = \mu_{m,c_+} - \mu_{m,c_-}. \tag{10}$$

We evaluate two contrasts: Deception minus Signaling and Bayesian minus Coordination. For a source model s and target model t, we add the source-model shift $v_s^{c_+,c_-}$ to representations from the

target model's c_- category. We then measure their maximum mean discrepancy (MMD) from the target model's c_+ representations, with lower MMD indicating greater distributional similarity.

The transferred shifts reduce MMD relative to a shuffled-label shift by 1.5–$1.6\times$, and relative to applying no shift by 1.7–$2.0\times$. Model-specific category shifts are also closely aligned with their pooled counterparts: cosine similarity is 0.96–0.98 across the two evaluated contrasts. The pooled direction outperforms an individual source-model direction in approximately 82% of source–target comparisons, consistent with pooling reducing estimation noise.

D.4 Interpretation

These results show that the evaluated model families organize the game suite along similar directions. They complement the dimensionality and reconstruction results in the main text: the models not only occupy a bounded response space, but also exhibit similar relative relationships among games within that space.

E Explained Variance Across PC Bands

E.1 Detailed Explained-Variance Analysis

Figure 5 summarizes the fraction of each model's variance explained by five PC bands for 20 representative models drawn from seven families. PCA is fitted on the full $n = 45$ model-mean embeddings; for a centered model-mean embedding $\mathbf{x}_m$ and PC band B, we report the normalized squared projection $V_m(B) = \|\Pi_B \mathbf{x}_m\|_2^2 / \|\mathbf{x}_m\|_2^2$. Thus, the percentages describe each model's explained-variance profile across PC bands.

E.2 Key Patterns Across Families

Model family is a weak proxy for representation geometry. Explained-variance profiles differ substantially even among sampled models with the same provider lineage. Among the Anthropic models in our sample, claude-opus-4.6 concentrates 56% of its explained variance in 1–2D, while claude-3.5-haiku places only 0.3% there and shifts 46% into the 7–12D band. Comparable within-family gaps appear for OpenAI, DeepSeek, and Qwen models. Thus, the relevant unit of analysis is the individual model rather than the provider family: alignment procedure, distillation, and scale can reshape the measured geometry as much as lineage does.

F Ablation Studies

F.1 Ablation: Embedding Model

Our main results use `text-embedding-3-small` (1,536D) (OpenAI, 2024). A legitimate concern is whether the dimensional hierarchy and composition results are artifacts of this specific text encoder. We therefore re-embed all 2,660 LLM responses with three alternative embedding models in Table 5, retrain the autoencoder at $d = 32$, and repeat the full analysis pipeline.

Table 5: Three alternative embedding models tested alongside the default choice.

Model	Provider	Dimensions	Type
text-embedding-3-small	OpenAI	1536	General
text-embedding-3-large	OpenAI	3072	General
E5-large-v2 (Wang et al., 2022)	Microsoft	1024	Contrastive
instructor-xl (Su et al., 2023)	HKU-NLP	768	Instruction-tuned

Key findings. The near agreement in model-means dimensionality supports the conclusion that compact between-model variation is not unique to the default text encoder. The difference in pooled-response dimensionality shows that absolute response-space dimension should not be interpreted as encoder-invariant.

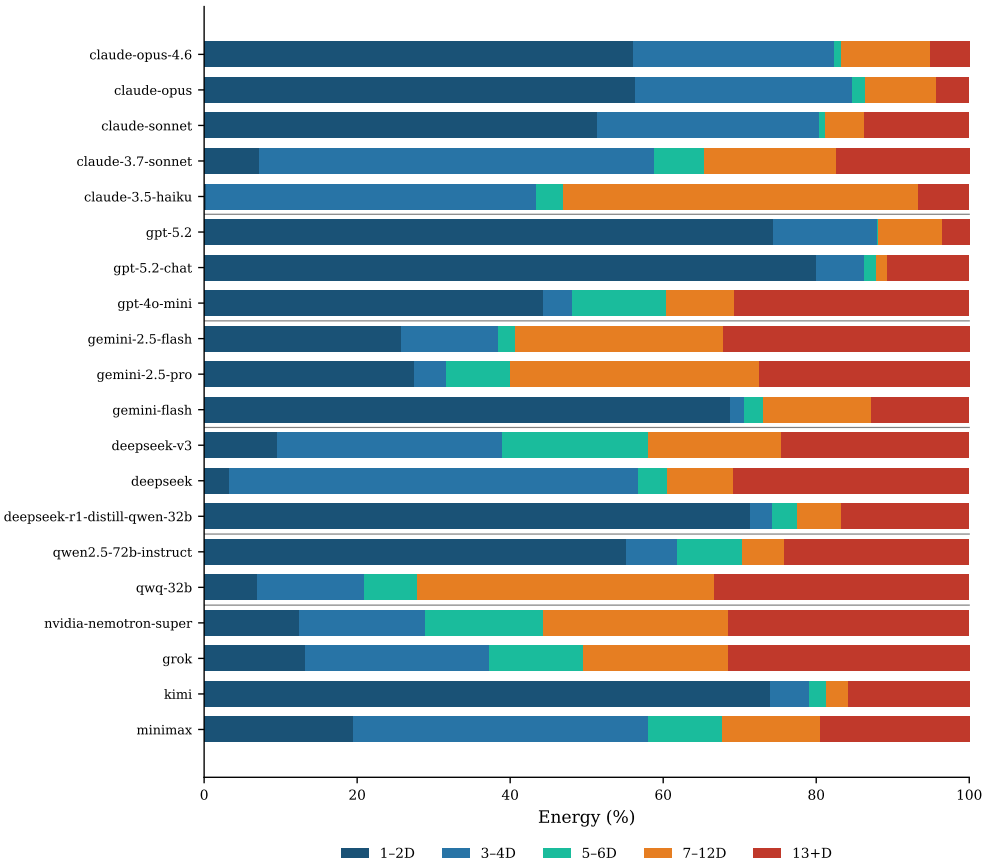

Figure 5: Explained-variance distribution across PC bands. Claude-3.5-haiku shifts 46% of explained variance to 7–12D while claude-opus-4.6 concentrates 56% in 1–2D; qwq-32b places substantially more mass in higher PC bands than qwen2.5-72b. Several generative models, including gpt-5.2-chat, kimi, and deepseek-r1-distill, place >70% of their explained variance in the first two PCs.

Table 6: **Comparison across response encoders.** Pooled-response dimensionality is encoder-dependent, while model-means dimensionality remains similar.

Measure	text-embedding-3-small	E5-large-v2
Model-means PCA d_{90}	16	17
Pooled-response PCA d_{90}	133	179
Primary ReLU plateau d^*	18	20

F.2 Ablation: Game Category

We test whether results depend on specific game categories by training autoencoders on a subset of games. The results are in Table 7. We observe that d_{90} drops by at most 1 compared to the full-game setting, indicating that the dimensional hierarchy is robust to game category removal.

Table 7: **Ablation: held-out game category.** Holding out one game category does not change the fundamental dimensionality structure.

Held-out category	None (full)	Signaling	Coordination	Bayesian	Deception
d_{90}	8	7	8	7	7

F.3 Ablation: Model Set

We evaluate whether the dimensional hierarchy is a stable property of strategic reasoning or an artifact of which models happen to be included. We run leave-one-out (LOO) validation for $n = 45$: for each model m_i, we hold it out entirely, retrain on the remaining models at $d = 32$ using identical hyperparameters, and recompute PCA on the retained-model embedding matrix. We observe that $d_{90} = 16$ in 44 of 45 folds, with the sole exception being Gemini 2.5 Pro ($d_{90} = 15$). No individual model, regardless of family, scale, or training lineage, drives the dimensional structure.

F.4 Ablation: Autoencoder architecture.

We test whether the estimated nonlinear plateau depends on the autoencoder architecture rather than the geometry of the response data. We evaluate 12 configurations spanning three widths (narrow, baseline, and wide), two activation functions (ReLU and GELU), and two weight-decay values (0 and 10^{-5}). For each configuration, we sweep bottleneck dimensions $d \in \{8, 12, 16, 18, 20, 24, 32\}$ and train five random seeds.

Among adequately expressive, unregularized configurations—the baseline and wide architectures without weight decay—the nonlinear reconstruction plateau remains within a narrow low-dimensional range. ReLU models yield a median plateau of $d^* = 18$, whereas GELU models yield $d^* = 24$. Increasing the network width beyond the baseline does not change the median plateau for either activation, and no adequate-capacity configuration produces $d^* > 24$.

The exact numerical value of d^* is therefore sensitive to the activation function and should not be interpreted as an architecture-invariant intrinsic dimension. Nevertheless, adequately expressive autoencoders consistently saturate within approximately 18–24 bottleneck dimensions. This supports the claim that the response space contains a bounded nonlinear reconstruction regime, rather than an exact universal core of 18 dimensions.

The remaining configurations serve as diagnostic controls. Narrow networks exhibit lower reconstruction quality and unstable plateaus, ranging from $d^* = 8$ for ReLU to $d^* = 32$ for GELU. Weight decay similarly suppresses reconstruction: five of the six regularized configurations have median $d^* = 8$, and several produce substantially reduced or negative EVR. These estimates reflect capacity-constrained or failed reconstruction rather than evidence for a smaller intrinsic geometry. We therefore distinguish robustness across adequately expressive models from diagnostic underfitting controls.

Table 8: Among adequately expressive configurations without weight decay, the median plateau remains between 18 and 24 dimensions. Narrow and regularized architectures are reported as diagnostic underfitting controls.

Width	Activation	Weight decay	Median d^*	d^* across seeds
Narrow	ReLU	0	8	12, 8, 8, 8, 8
Narrow	GELU	0	32	32, 32, 32, 32, 32
Baseline	ReLU	0	18	18, 20, 18, 18, 16
Wide	ReLU	0	18	20, 18, 24, 18, 18
Baseline	GELU	0	24	24, 24, 24, 24, 24
Wide	GELU	0	24	24, 24, 24, 24, 24
Narrow	ReLU	10^{-5}	8	18, 18, 8, 8, 8
Narrow	GELU	10^{-5}	8	8, 8, 8, 8, 8
Baseline	ReLU	10^{-5}	16	18, 18, 8, 8, 16
Baseline	GELU	10^{-5}	8	8, 8, 8, 8, 8
Wide	ReLU	10^{-5}	8	8, 16, 8, 8, 8
Wide	GELU	10^{-5}	8	8, 8, 8, 8, 8

G Robustness Evidence of Saturation

We evaluate whether the sublinear scaling pattern is robust to the choice of dimensionality diagnostic. Because different estimators characterize different aspects of the geometry, we do not expect them to agree in absolute scale. Instead, we test whether complementary global and local diagnostics exhibit the same qualitative saturation and held-out predictability as balanced PCA d_{90}. Both diagnostics are computed at each ecosystem size $n \in \{5, 10, 15, 20, 25, 30, 35, 40, 45\}$ using the same balanced subsampling protocol.

Participation ratio. As a threshold-free global spectral diagnostic, we compute the participation ratio $\mathrm{PR} = \frac{\left(\sum_i \lambda_i\right)^2}{\sum_i \lambda_i^2}$, where λ_i denotes the ith eigenvalue of the covariance matrix. PR measures the effective number of variance-carrying directions without selecting an explained-variance threshold. It reaches 20.7 ± 0.06 at $n = 45$, while its standard deviation across balanced subsets decreases from 1.2 at $n = 5$ to 0.06 at $n = 45$. Thus, the global spectral geometry becomes increasingly stable as the ecosystem grows.

Levina–Bickel MLE. As a complementary local-geometry diagnostic, we apply the Levina–Bickel k-nearest-neighbor intrinsic-dimension estimator (Levina & Bickel, 2004). Our primary estimate is the mean across $k \in \{8, 10, 12\}$; we additionally assess sensitivity over $k \in \{5, 8, 10, 12, 15\}$. The primary estimate reaches 14.5 ± 0.04 at $n = 45$. Although the estimator is less stable for the smallest ecosystems—its coefficient of variation is 24.0% at $n = 5$—the coefficient falls to 16.2% by $n = 15$ and remains below 15% for $n \geq 20$. The absolute estimate depends on neighborhood size (e.g., 18.7 for $k = 5$ at $n = 45$), but the flattening pattern is consistent across the tested values of k.

Held-out extrapolation. For each diagnostic, we fit a saturating exponential $f(n) = d_\infty \left(1 - ae^{-bn}\right)$ using only ecosystems with $n \leq 30$, and then predict the value at $n = 45$. The participation-ratio prediction differs from the observed value by $+0.5\%$, while the Levina–Bickel prediction differs by $+1.5\%$. Both therefore extrapolate to the full 45-model ecosystem within 2%. Together with the balanced-PCA result, these global-spectrum and local-geometry diagnostics show that the observed saturation is not specific to the d_{90} variance threshold or to a single notion of dimensionality.

H Behavior-Only Representation Analysis

The main analysis represents responses using text embeddings, so the observed geometry could partly reflect properties of the embedding model rather than strategic behavior. To test this possibility, we repeat the dimensional-scaling analysis after discarding the free-form reasoning and retaining only

Table 9: **Sensitivity of PCA dimensionality to the variance-explained threshold.** The ecosystem grows from 30 to 45 models. Model-means PCA measures between-model variation only.

Representation	Metric	$n = 30$	$n = 45$	Relative change
Model means	d_{85}	9	12	33.3%
	d_{90}	12	16	33.3%
	d_{95}	17	23	35.3%

each response's final observable action. This control asks whether bounded between-model structure remains detectable without a text encoder; it does not capture reasoning trajectories, uncertainty, or justification style.

Representation and preprocessing. We parse the final action from each response using game-specific templates applied to the final 300 characters of the response. Binary keyword matching, categorical one-hot encoding, and numeric extraction produce a 37-dimensional representation comprising 7 binary, 24 one-hot, and 6 normalized scalar features. Each response activates only the feature block associated with its game, making the representation highly sparse. We standardize each game-specific block using only responses from that game and then concatenate the blocks. Features with within-block standard deviation $\sigma < 10^{-3}$ are left unscaled. We avoid global z-scoring because it can amplify rare categories and destabilize the PCA spectrum.

Action parsing succeeds for 8,306 of 8,871 responses (92.3%). Parse rates are below 90% for Mafia voting (60.7%), prior updating in the first-price auction (76.9%), and buyer–seller negotiation (79.3%), and exceed 92% for all remaining games. We exclude responses with parsing failures.

Results. Our result is obtained by averaging features within each model and applying PCA to the resulting 45×37 matrix. The model-level representation yields $d_{90} = 13$, $d_{95} = 16$, and $\mathrm{PR} = 9.87$ at $n = 45$. The estimate is already similar at $n = 30$, where $d_{90} = 11.4 \pm 0.5$, indicating that the effective dimensionality stabilizes as additional models are included.

Response-level PCA provides a consistent result: d_{90} reaches 21 at $n = 35$, with zero variance across 10 repeats, and remains unchanged at $n = 40$ and $n = 45$; the participation ratio similarly saturates at $\mathrm{PR} = 22.0$. Because these response-level vectors are block-sparse, this result may partly reflect game-specific feature structure. We therefore treat the model-means analysis as the stronger test of between-model behavioral organization.

A supplementary autoencoder sweep detects a plateau at $d^* = 2$, but we do not interpret this value as behavioral intrinsic dimensionality: the autoencoder can compress the extremely sparse response-level representation nearly trivially. Taken together, the PCA results show that dimensional saturation persists after removing text embeddings, although they do not imply that the behavioral and embedding representations have identical geometry or dimensionality.

I Sensitivity to the Variance Threshold

The main analysis reports d_{90}, the number of principal components required to explain 90% of the variance. Because 90% is a conventional rather than theoretically privileged threshold, we test whether the dimensionality result depends on this choice. We recompute the endpoint dimensionalities at thresholds $\tau \in \{0.85, 0.90, 0.95\}$ for two complementary representations: PCA over all response embeddings, which measures the total response-space geometry, and PCA over model-level mean embeddings, which isolates between-model variation.

The saturation conclusion is robust to the threshold. For pooled-response PCA, increasing the ecosystem from 30 to 45 models changes the estimated dimensionality by only 1.4%, 3.1%, and 4.9% at the 85%, 90%, and 95% thresholds, respectively. Thus, the limited expansion observed at d_{90} is not specific to the 90% cutoff. The somewhat larger change at d_{95} indicates that additional models contribute more strongly to the low-variance spectral tail, but this contribution remains small relative to the 50% increase in the number of models. These endpoint comparisons are consistent with the flattening of the balanced-PCA scaling curve reported in Section 3.

Table 10: **Discussion-team compositions.** Team A uses three independently sampled agents instantiated from the same model. Teams B–D introduce increasing variation in model family, provider, and model type.

Team	Construction	Agent composition
A	Same model	GPT-5.2 $\times 3$
B	Same provider family	Claude Opus + Claude Sonnet + Claude Haiku
C	Cross-provider frontier	GPT-5.2 + Claude Opus + Gemini 2.5
D	Mixed closed/open models	GPT-5.2 + Llama 3.1 70B + Mistral 7B

The model-means analysis shows the same qualitative pattern. Between-model dimensionality increases by approximately 33%–35% across all three thresholds, less than the 50% increase in ecosystem size. The absolute dimension naturally increases as more of the spectral tail is retained, from $d_{85} = 12$ to $d_{95} = 23$ at $n = 45$, but the relative growth is nearly unchanged across thresholds. We therefore interpret the precise value of d_τ as threshold-dependent, while the substantive conclusion—that additional models yield diminishing expansion of the shared behavioral geometry—is threshold-robust.

J Multi-Agent Discussion

This section provides the complete team compositions and round-by-round results.

Protocol. We construct four teams, each containing three agents. The teams range from repeated instances of the same model to agents drawn from different model families and providers, as shown in Table 10. Each team discusses all 20 games for three rounds, with two independent trials per game. This produces 40 trials per team and 160 trials in total. In each round, every agent observes the responses from the preceding round and generates an updated response.

We embed the responses using `text-embedding-3-small` and compute two quantities:

1. *Within-team similarity*: the mean pairwise cosine similarity among the three agents' response embeddings. Higher values indicate greater agreement.
2. *Centroid distance*: the mean distance between the agents' response embeddings and the global centroid of the ecosystem responses. Changes in this quantity measure whether discussion moves responses systematically toward or away from the center of the observed response space.

Round-by-round results. Table 11 reports the complete trajectories. Within-team similarity increases across rounds for every team. The largest absolute increase occurs for Team A ($0.765 \rightarrow 0.866$), while the mixed closed/open team changes only slightly ($0.219 \rightarrow 0.239$). The same-provider Claude team begins with the highest similarity and changes from 0.837 to 0.858.

Centroid distance is comparatively stable. From the first to the final round, the largest absolute change is 0.026, observed for Team B. Multi-agent discussion therefore increases agreement within most teams without producing a comparably large systematic displacement relative to the global response centroid. This pattern is consistent with discussion primarily reorganizing responses within the observed behavioral geometry rather than substantially expanding it.

K Cross-Provider Sequential Composition

This section provides the complete protocol, model and game selection, and disaggregated results.

Protocol. Each trial consists of a three-turn proposal–critique–final workflow:

1. **Proposal.** The first model independently solves the game and produces a proposed action.
2. **Critique.** A model from a different provider reviews the proposal and recommends an action.

Table 11: **Multi-agent discussion trajectories.** Values are means across 40 trials per team. Pairwise cosine similarity measures within-team agreement; centroid distance measures displacement from the global ecosystem-response centroid.

	Pairwise cosine similarity			Centroid distance		
Team	Round 1	Round 2	Round 3	Round 1	Round 2	Round 3
A: Same model	0.765	0.837	0.866	0.962	0.947	0.942
B: Same provider family	0.837	0.848	0.858	0.770	0.748	0.744
C: Cross-provider frontier	0.566	0.631	0.645	0.839	0.826	0.825
D: Mixed closed/open	0.219	0.232	0.239	0.977	0.979	0.978

3. **Final decision.** The first model observes the original game, its proposal, and the critique before committing to a final action.

Because the first model both proposes and makes the final decision, we refer to it as the pipeline *anchor*. We name each directed protocol by its anchor and critic; for example, `anthropic_then_gemini` uses Anthropic as the anchor and Gemini as the critic.

Experimental design. We use one model from each of five providers: GPT-5.2, Claude Opus 4.6, Gemini 2.5 Pro, DeepSeek-V3, and Qwen 2.5 72B. We select five cross-provider pairs and evaluate each pair in both directions, producing ten directed orderings:

$$\begin{aligned} &\text{Anthropic} \leftrightarrow \text{DeepSeek}, \quad \text{Anthropic} \leftrightarrow \text{Gemini}, \\ &\text{Anthropic} \leftrightarrow \text{OpenAI}, \quad \text{OpenAI} \leftrightarrow \text{Qwen}, \\ &\text{DeepSeek} \leftrightarrow \text{Qwen}. \end{aligned}$$

The experiment covers five games spanning the four behavioral categories: credential signaling, stag hunt, first-price auction, bluffing poker, and negotiation. Running each of the ten orderings once on each game produces 50 transcripts and 150 model calls.

For each transcript, we record:

- whether the critic recommends a different action from the anchor's proposal;
- whether the anchor's final action differs from its proposal;
- whether reversing the provider order changes the final action; and
- whether reversing the provider order changes the strategic score.

Final actions are evaluated using the same game-specific criteria as in the main study. The bluffing-poker game admits a mixed-strategy solution without a single target action, so we report its valid-action rate rather than an equilibrium-action score.

Aggregate results. All 50 transcripts produced parseable, valid actions. Critics disagreed with the anchor's proposal in 12 of 50 transcripts (24%), while the anchor changed its final action in 9 of 50 transcripts (18%).

The ten directed orderings form 25 matched comparisons: five reciprocal provider pairs evaluated on five games. Reversing the anchor and critic changed the final action in 13 of these 25 comparisons (52%), but changed the strategic score in only 4 comparisons (16%). The four score-changing cases were:

- first-price auction, with Anthropic versus Gemini;
- negotiation, with Anthropic versus DeepSeek;
- credential signaling, with Anthropic versus Gemini; and
- credential signaling, with DeepSeek versus Qwen.

In these cases, the ordering anchored by Anthropic or DeepSeek received a score of 1, whereas the reverse ordering received a score of 0.

Table 12: **Composition outcomes by directed provider ordering.** Strategic score is averaged over the four games with a defined target action. Disagreement and proposal-to-final change rates are computed over all five games.

Anchor → critic	Score	Critic disagrees	Final changes
Anthropic → DeepSeek	1.00	0.00	0.00
Anthropic → Gemini	1.00	0.00	0.00
Anthropic → OpenAI	0.75	0.20	0.20
OpenAI → Anthropic	0.75	0.20	0.00
OpenAI → Qwen	0.75	0.00	0.00
Qwen → OpenAI	0.75	0.40	0.40
DeepSeek → Anthropic	0.75	0.60	0.60
DeepSeek → Qwen	0.72	0.00	0.00
Gemini → Anthropic	0.50	0.80	0.40
Qwen → DeepSeek	0.50	0.20	0.20

Table 13: **Composition outcomes by game.** Strategic score and proposal-to-final change rate are averaged over the ten directed provider orderings. The value reported for bluffing poker is its valid-action rate.

Game	Category	Score	Final changes
B1: First-price auction	Bayesian	0.89	0.30
D3: Negotiation	Deception	0.90	0.20
S2: Credential signaling	Signaling	0.80	0.00
C3: Stag hunt	Coordination	0.40	0.30
D1: Bluffing poker	Deception	1.00[†]	0.10

[†]Valid-action rate; the game has no single target action.

Interpretation. Cross-provider critiques infrequently change the anchor's proposal. Provider order has a larger effect on which action is selected than on whether that action satisfies the game-specific scoring criterion. These results indicate that the proposing-and-deciding model remains an important determinant of the pipeline outcome, even when its proposal is reviewed by a model from another provider.

L Cross-Domain Validation on Open-Ended Generation

The main experiments establish a bounded nonlinear core for strategic reasoning. Here we test whether this structure transfers to a substantially different response domain. We use open-ended generation as a boundary check because it allows greater variation in content, style, and response format than the structured strategic-game prompts.

Experimental design. We select 50 queries spanning seven categories from the taxonomy of Jiang et al. (2025): creative writing, reasoning, knowledge, advice, coding, open-ended questions, and mathematics. We collect responses from the same 45-model ecosystem used in the main experiments, obtaining 2,242 valid responses after filtering unsuccessful requests.

We embed the responses with `text-embedding-3-small` and repeat the reconstruction analysis using the same primary ReLU autoencoder architecture and plateau-selection criterion as in Section 4. We evaluate validation EVR across bottleneck widths $d \in \{8, 12, 16, 18, 20, 24, 28, 32\}$. We also compute model-means PCA and full pooled PCA using the definitions in Section 2.

Results. The open-ended corpus reaches a nonlinear plateau at $d^* = 18$, the same value obtained for strategic games under the primary ReLU protocol. Its model-means PCA dimensionality is also

Table 14: **Cross-domain comparison of strategic games and open-ended generation.** The estimated nonlinear plateau and between-model dimensionality remain nearly unchanged, although the full response space is broader for open-ended generation.

Measure	Strategic games	Open-ended generation
Nonlinear plateau d^*, primary ReLU protocol	18	18
Model-means PCA d_{90}	16	17
Full pooled PCA d_{90}	133	154

nearly unchanged: $d_{90} = 17$ for open-ended generation, compared with $d_{90} = 16$ for strategic games. Thus, changing the task domain does not materially expand the estimated between-model core.

The full pooled PCA result provides an important contrast. Open-ended generation has a higher response-level dimensionality, $d_{90} = 154$, compared with 133 for strategic games. The open-ended prompts therefore produce a broader space of individual responses even though the dimensionality of the differences between models remains nearly constant. This dissociation indicates that the cross-domain agreement is not caused by the two datasets having identical response variation.