RELATIVE CAPABILITY ORGANIZES THE GEOMETRY OF ANSWER DISTRIBUTIONS IN TEST-TIME SCALING

Anonymous authors
Paper under double-blind review

ABSTRACT

Test-time scaling naturally produces distributions over candidate answers, revealing how a model's responses vary across repeated samples on the same problem. While prior work has connected individual properties of these distributions to difficulty, uncertainty, and aggregation performance, whether the distribution as a whole exhibits a shared structure across models and tasks remains unclear. We find that ranked answer distributions exhibit a shared geometry beyond ranking and probability constraints, systematically organized by model–problem relative capability. As relative capability increases, distributions progress from fragmentation, through structured competition among leading answers, to single-cluster dominance. Correct-answer location is reorganized within this geometry: it shifts sharply toward the leading cluster in the intermediate regime, after which answer dominance increasingly coincides with correctness. This empirical law generalizes to held-out models and tasks and extends to constrained answer spaces through a systematic geometric deformation. Together, these findings offer a new structural perspective on test-time scaling, revealing systematic organization in LLM reasoning behavior across models and problems.

1 INTRODUCTION

Test-time scaling (TTS) (OpenAI, 2024; Snell et al., 2025) has become an important paradigm for improving the reasoning performance of large language models (LLMs) by allocating additional computation at inference time. A widely used approach is to sample multiple reasoning trajectories or candidate answers and aggregate the resulting outputs (Brown et al., 2024). Methods such as self-consistency (SC) exploit information contained in the resulting answer distribution, selecting answers according to their frequency or agreement across repeated samples (Wang et al., 2023). The answer distribution therefore provides a natural representation of how a model behaves under repeated inference and plays a central role in sampling-based test-time reasoning.

Prior work has studied many important properties of these distributions, including answer entropy and diversity, agreement among samples, competition between leading answers, and the sample complexity of aggregation (Farquhar et al., 2024; Feng et al., 2025). These studies have also revealed systematic relationships between such statistics and problem difficulty: harder problems generally produce more diverse and less concentrated answers, while stronger agreement often allows reliable decisions with fewer samples (Yeom et al., 2026; Huang et al., 2026). However, existing analyses largely characterize answer distributions through individual summary statistics or particular aggregation behaviors. Much less is known about whether the *full answer distribution* itself follows a systematic structure across models and tasks, and especially how this structure changes with difficulty. Characterizing such regularity could provide a unified view of how reasoning uncertainty evolves with model capability and help explain when sampling-based aggregation succeeds or fails.

In this work, we conduct a large-scale empirical study of answer distributions across diverse LLMs and reasoning benchmarks. We first find that ranked answer distributions exhibit a non-trivial shared geometry that cannot be explained solely by ranking, normalization, and finite-sampling constraints. We then use the Rasch model (Rasch, 1960; Ding et al., 2024) to define model–problem relative capability and find that this single coordinate systematically organizes the shared geometry across models and benchmarks. As relative capability increases, distributions move from fragmentation

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.

to structured competition and ultimately dominance, while the correct answer simultaneously shifts toward the leading cluster.

We further show that this relationship generalizes across held-out models and benchmarks and remains predictive under broader model–task shifts. For constrained answer spaces, the same underlying law largely persists after accounting for the finite answer set, with constrained distributions well approximated by projections of the corresponding open-answer distributions.

Our main contributions are as follows:

- We identify a non-trivial shared geometry of ranked answer distributions beyond the structure induced by basic distributional constraints.
- This geometry is organized by a 1D problem-relative-capability coordinate, along which answer distributions evolve from fragmentation to dominance, with correctness shifting to the leading cluster.
- We establish transfer across held-out models and benchmarks and characterize finite answer spaces as systematic deformations of the open-answer law.

2 Previous Studies of Answer Distributions in Test-Time Scaling

Repeated reasoning samples were initially exploited through self-consistency, which aggregates answers from diverse reasoning paths rather than relying on a single decoding trajectory (Wang et al., 2023). Subsequent work has made this distributional view increasingly explicit, treating repeated samples as an empirical approximation to the model's underlying answer distribution (Li et al., 2024). A growing body of work has shown that this distribution varies systematically with problem difficulty: higher difficulty is generally associated with greater answer entropy and diversity, while more concentrated distributions can often be resolved with fewer samples, motivating difficulty- and entropy-adaptive sampling strategies (Yeom et al., 2026). Other studies have focused on competition among leading answers, showing that self-consistency depends on whether the correct answer is the population mode and that a smaller gap between the strongest candidates substantially increases the samples required for stable aggregation (Feng et al., 2025; Huang et al., 2026). Recent work further shows that strong agreement or a dominant answer does not necessarily imply correctness, reinforcing the close connection between answer-distribution properties and aggregation behavior (Chen et al., 2024; Cai et al., 2026).

Despite this progress, prior studies have primarily characterized answer distributions through summary statistics such as entropy, diversity, majority strength, and the gap between leading candidates, thereby compressing much of the information in the full distribution. It therefore remains unclear whether full answer distributions exhibit reproducible structure, and whether phenomena such as fragmentation, concentration, and leading-answer competition reflect a common distributional organization. It is also unknown whether such structure is shared across models and reasoning tasks. We therefore begin by characterizing the geometry of the full answer distribution itself.

3 Characterizing Answer-Distribution Geometry

We study seven instruction-tuned language models: Llama3.1-8B (Grattafiori et al., 2024), Qwen2.5-7B (Yang et al., 2024), Qwen3-8B (Yang et al., 2025), Mistral-7B (Jiang et al., 2023), OLMo-3-7B (Olmo et al., 2026), InternLM3-8B (InternLM Team, 2025), and Qwen3.5-9B (Qwen Team, 2026a), evaluated on seven open-answer mathematical and scientific reasoning benchmarks: AIME-2024 (Zhang & Math-AI, 2024), AIME-2025 (Zhang & Team, 2025), AMC-2023 (Math-AI, 2024), MATH500 (Hendrycks et al., 2021), Minerva-Math (Lewkowycz et al., 2022), OlympiadBench (He et al., 2024), and OlympiadPhysics (He et al., 2024), comprising 1,782 unique problems in total.

For each available model–problem pair, we independently sample 128 reasoning trajectories and merge trajectories with equivalent final answers into answer clusters. We then sort the clusters by empirical mass, yielding a ranked distribution $p_1 \geq p_2 \geq \cdots$. Ranking removes answer identity and places distributions from different problems in a common coordinate system, where the same

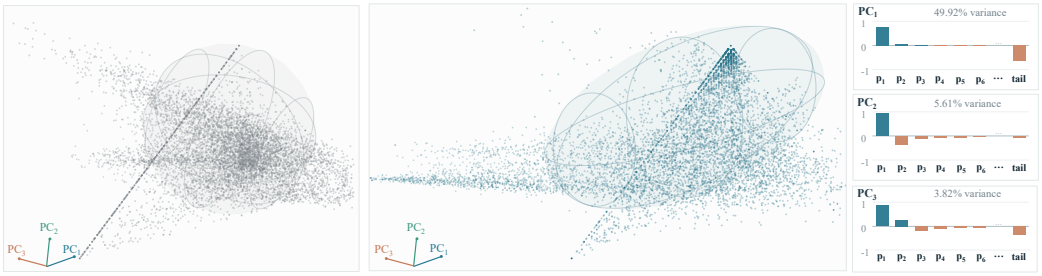


Figure 1: **Shared geometry of ranked answer distributions.** (A) Matched null distributions preserving ranked-simplex, support-size, and finite-sampling constraints. (B) Empirical distributions exhibit additional structure, particularly along the dominant PC1 direction. (C) PCA loadings show that PC1 mainly captures dispersion versus concentration, while PC2–PC3 characterize competition among leading clusters.

dimension always denotes the j-th most frequent answer. We represent each distribution as

$$D = (p_1, p_2, \ldots, p_{20}, p_{\text{tail}}), \qquad p_{\text{tail}} = 1 - \sum_{j=1}^{20} p_j,$$

using the leading 20 clusters together with an aggregated tail to retain the dominant distributional structure in a fixed-dimensional representation (see Appendix A for the handling of length-truncated and unresolved generations).

3.1 Ranked Answer Distributions Reveal a Shared Geometry

We first apply PCA to the ranked distributions to examine their shared geometric structure. Because ranked probability vectors are constrained by

$$p_1 \geq p_2 \geq \cdots \geq 0, \qquad \sum_j p_j = 1,$$

part of the resulting geometry may arise from ranking and simplex constraints alone. We therefore construct a matched null distribution for each case using a symmetric Dirichlet–multinomial model, with the dimension matched to the observed number of answer clusters K and the multinomial draw count matched to the number of valid samples N. The null cluster masses are ranked and transformed into the same representation as the empirical distributions, preserving case-level support size and finite-sampling effects.

The left side of Figure 1 compares the null and empirical PCA geometries under the same view. The null distributions already form a characteristic triangular plane, showing that part of the geometry is induced by the ranked-simplex constraints. The empirical distributions, however, differ systematically from this reference: they extend substantially further along the dominant PC1 direction and show a pronounced concentration near the upper vertex of the triangular plane, corresponding to distributions dominated by a single leading cluster.

The PCA loadings shown on the right side of Figure 1 clarify these directions. PC1 primarily captures the transition between dispersed and concentrated distributions, whereas PC2 and PC3 describe how probability mass is allocated among the leading clusters. Together, these results show that ranked answer distributions exhibit a shared empirical geometry beyond that induced by ranking and probability constraints.

3.2 Model–Problem Relative Capability

To study how difficulty relates to the observed geometry, we need a difficulty measure that accounts for model capability. A problem does not have the same effective difficulty for every model: the

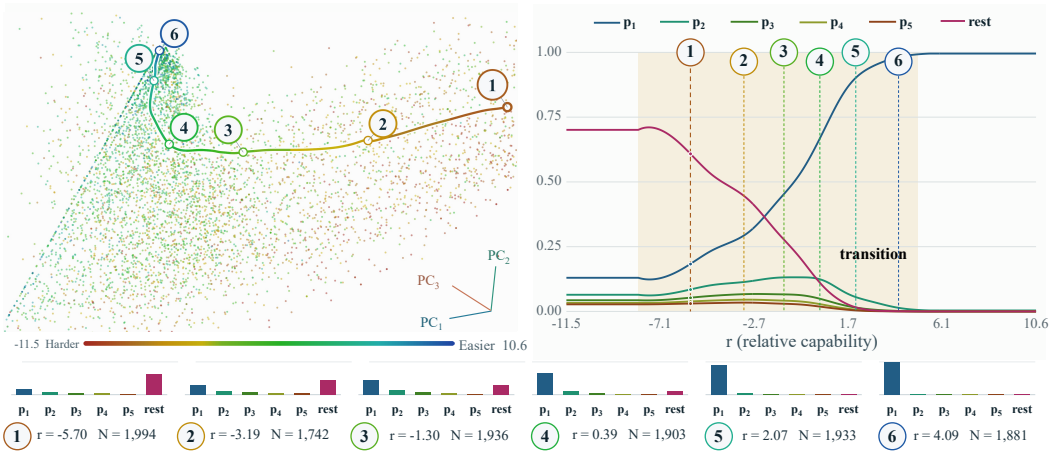


Figure 2: **Relative capability organizes answer-distribution geometry.** (A) Distribution centroids trace an ordered trajectory in PCA space as relative capability r increases. (B) The fitted law $F(r)$ shows how ranked probability mass varies along the same coordinate.

same item may be easy for a stronger model but difficult for a weaker one. We therefore use a Rasch model to characterize their relative capability:

$$P(y_{mi} = 1) = \sigma(\theta_m - \beta_i) = \sigma(r_{mi}), \qquad r_{mi} \equiv \theta_m - \beta_i,$$

where θ_m is the capability of model m, β_i is the difficulty of item i, and $\sigma(\cdot)$ is the logistic function. Thus, r_{mi} can be estimated from the model's probability of answering a particular problem correctly, with $r = 0$ corresponding to a 50% correctness probability.

Because r is estimated from correctness observations, directly using a case's own correctness in its assigned coordinate would introduce a mechanical dependence in subsequent analyses. We therefore use cross-fitted estimates, excluding the focal observation when estimating its relative capability (see Appendix B.3). This yields a common model–problem scale on which the relationship between difficulty and answer-distribution geometry can be examined.

4 Relative Capability Organizes Answer-Distribution Geometry

Having established that ranked answer distributions exhibit a non-trivial shared geometry, and introduced r_{mi} as a common measure of model–problem relative capability, we next investigate how relative capability is related to this geometry. We first examine how answer distributions move through the shared geometric space as r_{mi} varies, and then study how correctness changes along the same coordinate.

4.1 Relative Capability Organizes the Geometry

We compute cross-fitted relative capability r_{mi} for all model–problem pairs across seven models and seven benchmarks using the Rasch model introduced in Section 3.2. We group samples by r_{mi}, compute the centroid of each group in the PCA space, and connect these centroids to obtain a trajectory along the relative-capability axis.

As shown in Figure 2A, this trajectory forms a continuous path through the shared geometry. At low r_{mi}, the centroids lie away from the triangular region identified in Section 3. As r_{mi} increases, the trajectory first moves mainly along the leading principal direction toward this region, and then progressively toward its apex, corresponding to distributions increasingly dominated by a single answer cluster. Although the PCA representation is learned without difficulty information, relative capability induces a highly ordered progression through the space.

We further examine this progression directly in the ranked-distribution space by fitting each component of the full ranked representation $D = (p_1, \ldots, p_{20}, p_{\text{tail}})$ as a function of r. We denote the

resulting empirical law as

$$\hat{D}(r) = F(r),$$

where F maps relative capability to the expected ranked answer distribution. For visualization in Figure 2B, we show the leading five clusters explicitly and aggregate all remaining probability mass into a single *rest* category.

As an in-distribution predictive check within the development pool, we perform case-level cross-validation: we fit $F(r)$ on the training cases and use it to predict the ranked distributions of held-out cases. We compare it with a constant baseline that assigns every held-out case the mean distribution of the corresponding training data.

Conditioning on r_{mi} substantially improves shape prediction. On the held-out fold, the mean L_1 distance between predicted and observed distributions decreases from 0.793 for the global-mean baseline to 0.453 for $F(r)$, a 42.8% reduction in prediction error. The same pattern holds in PCA space, where the mean Euclidean error decreases from 0.421 to 0.230. These results indicate that the shared answer-distribution geometry is strongly organized by the one-dimensional relative-capability coordinate r_{mi}.

4.2 Reorganization of Probability Mass along Relative Capability

As shown in Figure 2B, this process cannot be described simply as increasing concentration with capability. At low r, probability mass is dispersed across many answer clusters: repeated reasoning produces diverse answers without a stable leading candidate. The distribution then enters a pronounced transition region, where mass concentrates into a small number of leading clusters. A few candidate answers now recur consistently, but none yet dominates, giving rise to structured competition. At high r, this competition is progressively resolved as one cluster absorbs most of the probability mass and becomes dominant. The overall progression can therefore be summarized as

$$\boxed{\text{fragmentation} \longrightarrow \text{structured competition} \longrightarrow \text{dominance.}}$$

We further examine how probability mass is redistributed within the transition region. The fitted component curves in Figure 3A suggest two forms of redistribution: mass transfer from the tail to the leading clusters, and redistribution within the leading clusters toward the top-ranked cluster. To characterize their relative strength, we examine the local derivatives of the fitted curves along r.

In *Stage I*, tail-to-head redistribution is stronger: tail mass decreases while several leading clusters simultaneously gain probability mass. Around $r \approx -0.6$, the relative strength of the two components reverses. In *Stage II*, redistribution within the head becomes dominant: the top-ranked cluster continues to gain mass, while the other leading clusters cease to grow and gradually decline. The two processes overlap around this point rather than forming a sharp boundary.

The transition therefore reflects a change in the form of mass reorganization: diffuse responses are first organized into a small set of recurring candidates, after which competition among these candidates is progressively consolidated toward the dominant cluster.

4.3 Correctness Transition and Self-Consistency

We now restore correctness information and examine how the position of the correct-answer cluster changes with relative capability. Figure 3B reports the probability that the correct answer is not sampled, or appears at ranks p_1 through p_5, with all lower ranks grouped as *rest*. As r increases, the probability of not sampling the correct answer steadily decreases, while correctness moves from lower-ranked clusters toward the leading positions and eventually concentrates at p_1.

Correctness therefore undergoes a pronounced transition along r, but over a narrower range than the distributional transition. The shape begins to reorganize first, while the migration of correctness toward the leading cluster occurs later. The two transitions are thus coupled but not synchronous: the distribution first develops a small set of stable competitors, after which the correct answer increasingly becomes the dominant one.

This perspective also provides a useful decomposition of self-consistency. Its success requires two conditions: the correct answer must first be sampled (*reachability*), and, once sampled, it must

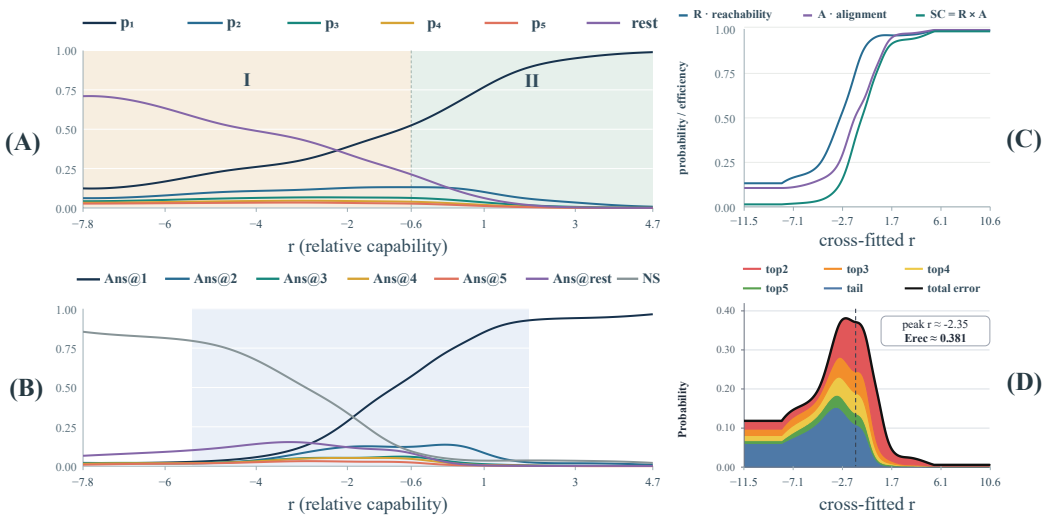


Figure 3: **Distributional and correctness transitions along relative capability.** (A) Probability-mass redistribution shifts from tail-to-head transfer to concentration toward p_1. (B) The correct answer progressively moves toward the leading cluster. (C) Alignment lags behind reachability. (D) Recoverable errors peak at intermediate r and are increasingly concentrated among leading competitors.

become the largest cluster (*alignment*). We define

$$R(r) = P(\text{correct sampled} \mid r), \qquad A(r) = P(p_1 \text{ correct} \mid \text{correct sampled}, r).$$

As shown in Figure 3C, alignment lags behind reachability as r increases. This creates an intermediate region in which the correct answer is already present among the samples but self-consistency still fails to select it.

We refer to these cases as *recoverable errors*, whose probability is

$$E_{\text{rec}}(r) = R(r)\,[1 - A(r)].$$

Figure 3D shows that recoverable error first increases with r, peaks around $r \approx -2.35$, and then declines rapidly. Around this point, the gap between reachability and alignment is largest: the model can already generate the correct answer, but frequency-based aggregation does not yet reliably promote it to the largest cluster. In terms of error composition, as relative capability increases, recoverable errors progressively shift toward higher-ranked competing clusters and, on the higher-r side, become concentrated in cases where self-consistency selects the wrong answer between the top two clusters.

These results suggest that the success of self-consistency is largely driven by the progressive alignment between answer dominance and correctness as relative capability increases. Its remaining recoverable failures are concentrated at intermediate capability, where the key difficulty is no longer generating the correct answer, but resolving competition among a few leading candidates—especially the top two.

5 Generalization and Transferability

Having established the predictive relationship between relative capability and answer-distribution shape within the development pool, we next ask whether this relationship transfers beyond the observations used to establish it. We retain the same prediction setup, fitting $F_{\text{train}}(r)$ on the designated training data and evaluating its predictions on held-out test cases using Jensen–Shannon divergence (JSD).

Table 1: **Cross-identity transfer of the relative-capability law.** Each model or benchmark is entirely held out from law fitting. Gain denotes the JSD reduction relative to the training-mean baseline.

Model	JSD ↓	Gain ↑	Benchmark	JSD ↓	Gain ↑
Llama3.1-8B	0.0742	46.5%	AIME-2024	0.0827	35.4%
Mistral-7B	0.0781	59.9%	AIME-2025	0.0855	36.9%
OLMo-3-7B	0.1069	38.1%	AMC-2023	0.0722	45.7%
Qwen2.5-7B	0.0691	44.7%	MATH500	0.0488	69.8%
Qwen3-8B	0.0576	57.7%	Minerva	0.0900	38.8%
Qwen3.5-9B	0.0542	70.9%	OlympiadBench	0.0864	36.5%
InternLM3-8B	0.0691	45.4%	OlympiadPhysics	0.0936	45.7%

As a baseline, we use the mean training distribution $\bar{D}_{\text{train}} = \frac{1}{N_{\text{train}}} \sum_{i \in \mathcal{D}_{\text{train}}} D_i$, which ignores relative capability. We report the relative reduction in prediction error,

$$G = 1 - \frac{\mathbb{E}[\text{JSD}(D, F_{\text{train}}(r))]}{\mathbb{E}[\text{JSD}(D, \bar{D}_{\text{train}})]}.$$

Lower JSD and higher G indicate stronger transfer. We consider two progressively stronger questions:

- **Can the law transfer to an entire model or benchmark identity that is absent from the training data?**
- **Can a law fitted on the original development pool remain predictive under new external tasks and substantially larger models?**

5.1 Generalization across Model and Benchmark Identities

Observations from the same model may share model-specific generation characteristics, while problems from the same benchmark may share task structure and domain properties. To test whether the empirical law extends beyond such identity-specific regularities, we remove one entire model or benchmark, fit the law using all remaining identities, and evaluate the frozen law exclusively on the held-out identity.

Table 1 reports the results. The frozen law improves over the training-mean baseline for every held-out model and benchmark. Across the seven models, gains range from 38.1% to 70.9%, while across the seven benchmarks they range from 35.4% to 69.8%.

The consistently positive gains show that the $r \to D$ relationship is not tied to the particular model or benchmark identities used to estimate it. Instead, relative capability provides a transferable coordinate for organizing answer-distribution geometry across both models and tasks.

5.2 Generalization beyond the Development Pool

Although the previous experiment holds out complete identities, all models and benchmarks still come from the same development pool and may share broader properties of that pool. We therefore adopt a stricter setting by freezing $F_{\text{core}}(r)$, fitted once on the original core data, and evaluating it on HMMT25 (Dekoninck et al., 2026), a new and harder competition-mathematics benchmark. We evaluate four model families already represented in the development pool together with two additional models, Qwen3.6-35B-A3B (Qwen Team, 2026b) and DeepSeek-V4-Flash (DeepSeek-AI, 2026), without using any HMMT25 answer-distribution observations to fit

Table 2: Generalization of the frozen core law beyond the development pool.

HMMT25			
Model	Scale	JSD ↓	Gain ↑
Qwen2.5	7B	0.0907	-1.0%
InternLM3	8B	0.0768	17.8%
Qwen3	8B	0.0936	18.2%
Qwen3.5	9B	0.1756	-2.8%
Qwen3.6	35B	0.0739	47.1%
DS-v4-Flash	284B	0.0698	39.2%

or smooth $F_{\text{core}}(r)$. This setting therefore tests transfer to a new benchmark and, through the two additional models, to substantially larger total parameter counts.

As shown in Table 2, transfer becomes less uniform under this stronger distribution shift, with four of the six model–benchmark conditions retaining positive gains over the training-mean baseline. Notably, the frozen law remains strongly predictive for the two newly introduced models, achieving gains of 47.1% for Qwen3.6-35B-A3B and 39.2% for DeepSeek-V4-Flash.

The small negative gains for Qwen2.5 and Qwen3.5 indicate that relative capability does not fully determine answer-distribution geometry under every external condition. Nevertheless, its transfer to unseen identities and its continued predictive power under substantially shifted model–task conditions provide strong evidence that relative capability captures a shared organizing structure of answer-distribution geometry rather than a pattern specific to the original models, benchmarks, or development pool.

6 Answer-Space Constraints Deform the Open Law

The empirical law established above is derived primarily from open-answer reasoning tasks. However, multiple-choice benchmarks such as GPQA-Diamond exhibit noticeably different distributional behavior and are less well predicted by the open-answer law. A key difference is that their answer space is explicitly constrained: probability mass is restricted to only a few candidate answers, making highly fragmented distributions difficult to realize. Similar constraints also arise in practical decision-making and action-selection problems. We therefore ask whether finite answer spaces induce a different distributional law, or instead systematically deform the open-answer law.

To isolate this effect, we construct controlled constrained variants of Minerva and OlympiadPhysics. For each model–problem pair, we take the k most frequent answers under open sampling as candidate options and construct k-choice versions for $k = 3, \ldots, 7$, while retaining the original open-answer case as a paired reference. This allows us to vary the presence and strength of answer-space constraints while holding the model and problem fixed. We first study these controlled pairs and then examine whether the observed relationship extends to naturally constrained benchmarks, including GPQA-Diamond (Rein et al., 2024) and the more challenging SuperGPQA (M-A-P & Seed, 2025).

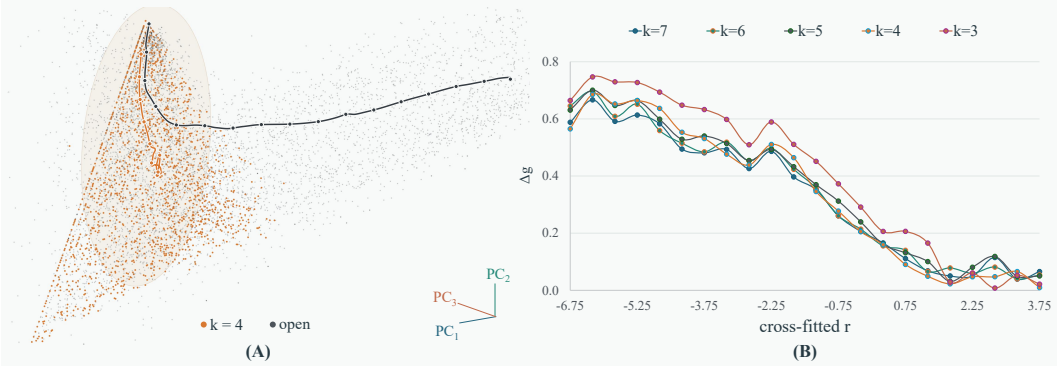


Figure 4: **Finite answer spaces deform the open-answer geometry.** (A) Constrained distributions are projected toward the triangular region and progressively approach the open trajectory as r increases. (B) The geometric deviation Δg decreases with relative capability across different answer-space sizes.

6.1 Finite Answer Spaces Act Primarily as a Geometric Projection

We first examine how answer-space constraints alter the distribution geometry. Across all controlled settings with $k = 3, \ldots, 7$, as well as the naturally constrained GPQA-Diamond and SuperGPQA, the distributions concentrate on the triangular plane in PCA space. Figure 4A illustrates this pattern using $k = 4$ as a representative case. Geometrically, the change resembles a redistribution of probability mass from the fragmented directions of the open space into the leading answer clusters.

Motivated by this observation, we define a simple top-k projection. For an open ranked distribution $D^{\text{open}} = (p_1, p_2, \ldots)$, we retain the first k clusters, remove the remaining mass, and renormalize:

$$P_k(D^{\text{open}}) = \frac{(p_1, \ldots, p_k)}{\sum_{j=1}^{k} p_j}. \tag{1}$$

We then apply the same transformation to the fitted open law $F(r)$. The resulting $P_k(F(r))$ uses only the known answer-space size k, without fitting a separate constrained law. As shown in Table 3, this simple projection substantially reduces prediction error on the constrained Minerva and OlympiadPhysics variants and also improves the characterization of GPQA-Diamond and SuperGPQA.

Table 3: Effect of finite-space projection on constrained-law prediction, measured by Jensen–Shannon divergence (JSD).

Dataset	k	Open	Proj.	Gain
Minerva	3–7	0.1278	0.0152	88.1%
OlympiadPhysics	3–7	0.1580	0.0190	88.0%
GPQA-Diamond	4	0.1184	0.0110	90.7%
SuperGPQA	4	0.1341	0.0283	78.9%
SuperGPQA	10	0.0828	0.0380	54.2%

These results suggest that a substantial part of the deviation from the open law can be accounted for by answer-space geometry. For a naturally constrained task with known k, redistributing the mass outside the available answer set provides a simple way to adapt the open-law prediction at the same relative capability.

6.2 Constrained Distributions Progressively Rejoin the Open Law

Another regularity Figure 4A shows is that constrained distributions remain near the triangular plane, while the open trajectory progressively approaches the same region as relative capability increases. We quantify this convergence using Δg, the distance between the constrained centroid and the corresponding open-law centroid at the same r. As shown in Figure 4B, Δg decreases consistently across $k = 3, \ldots, 7$, indicating that the effect of answer-space constraints is strongest at low relative capability and weakens as r increases.

This trend is consistent with the projection view: low-r open distributions place substantial probability mass outside the leading k clusters, whereas increasing r naturally concentrates more mass within them. Consequently, progressively less mass is affected by the finite answer space, and the constrained distribution approaches the open law.

However, the low-r deformation is not fully explained by mass redistribution alone. We therefore examine

$$\Delta p_1(r) = p_1^{\text{con}} - p_1^{\text{proj}},$$

where p_1^{proj} is obtained by projecting the paired open distribution. The residual is dominated by additional top1 concentration: constrained generation assigns more probability mass to the leading cluster than predicted by projection alone. The strength of this effect varies substantially across settings, and its underlying mechanism remains unclear.

Overall, finite answer spaces primarily deform the open law through top-k projection, with the deformation strongest at low relative capability and progressively diminishing as r increases. A smaller residual effect remains in the form of additional top1 concentration.

7 Conclusion

We show that ranked answer distributions exhibit a shared geometry systematically organized by model–problem relative capability. This coordinate jointly structures distributional shape and correctness, transfers across held-out models and benchmarks, and remains informative under broader model–task shifts. Finite answer spaces largely preserve this organization through a systematic projection of the open-answer law. Together, these findings reveal a previously underexplored regularity in repeated LLM reasoning and provide a foundation for understanding, modeling, and ultimately improving test-time scaling through the structure of answer distributions.

AI USE STATEMENT

In this work, we used generative AI tools to aid and polish the writing of the manuscript, to assist with literature retrieval and discovery, and to support research ideation and the implementation and execution of analysis experiments. All AI-assisted writing was reviewed and revised by the authors. Retrieved literature was manually checked against the original sources before citation. For AI-assisted analyses, the authors inspected the underlying data, reviewed the generated or modified code, and verified the resulting outputs and conclusions. We take responsibility for the final content of this work, including all text, claims, analyses, and artifacts produced with the aid of generative AI.

REFERENCES

Bradley C. A. Brown, Jordan Juravsky, Ryan Ehrlich, Ronald Clark, Quoc V. Le, Christopher Ré, and Azalia Mirhoseini. Large language monkeys: Scaling inference compute with repeated sampling. *CoRR*, abs/2407.21787, 2024. doi: 10.48550/ARXIV.2407.21787. URL https://doi.org/10.48550/arXiv.2407.21787.

Meng Cai, Lars Kulik, and Farhana Choudhury. ARBITER: reasoning trajectory basins and majority vote failures in test-time sampling. *CoRR*, abs/2605.26172, 2026. doi: 10.48550/ARXIV.2605.26172. URL https://doi.org/10.48550/arXiv.2605.26172.

Lingjiao Chen, Jared Quincy Davis, Boris Hanin, Peter Bailis, Ion Stoica, Matei A. Zaharia, and James Y. Zou. Are more LLM calls all you need? towards the scaling properties of compound AI systems. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/51173cf34c5faac9796a47dc2fdd3a71-Abstract-Conference.html.

DeepSeek-AI. Deepseek-v4: Towards highly efficient million-token context intelligence. *CoRR*, abs/2606.19348, 2026. doi: 10.48550/ARXIV.2606.19348. URL https://doi.org/10.48550/arXiv.2606.19348.

Jasper Dekoninck, Nikola Jovanovic, Tim Gehrunger, Kári Rögnvaldsson, Ivo Petrov, Chenhao Sun, and Martin T. Vechev. Beyond benchmarks: Matharena as an evaluation platform for mathematics with llms. *CoRR*, abs/2605.00674, 2026. doi: 10.48550/ARXIV.2605.00674. URL https://doi.org/10.48550/arXiv.2605.00674.

Mucong Ding, Chenghao Deng, Jocelyn Choo, Zichu Wu, Aakriti Agrawal, Avi Schwarzschild, Tianyi Zhou, Tom Goldstein, John Langford, Animashree Anandkumar, and Furong Huang. Easy2hard-bench: Standardized difficulty labels for profiling LLM performance and generalization. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/4e6f223052759665139 90f53cec908e0-Abstract-Datasets_and_Benchmarks_Track.html.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nat.*, 630(8017):625–630, 2024. doi: 10.1038/S41586-024-07421-0. URL https://doi.org/10.1038/s41586-024-07421-0.

Austin Feng, Marius Alonso, and Ambroise Odonnat. Optimal self-consistency for efficient reasoning with large language models. *CoRR*, abs/2511.12309, 2025. doi: 10.48550/ARXIV.2511.12309. URL https://doi.org/10.48550/arXiv.2511.12309.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, et al. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, et al. Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3828–3850, 2024.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

Baihe Huang, Shanda Li, Tianhao Wu, Yiming Yang, Ameet Talwalkar, Kannan Ramchandran, Michael Jordan, and Jiantao Jiao. Sample complexity and representation ability of test-time scaling paradigms. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 54262–54298, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/592729d610953abb6d16065773316d00-Paper-Conference.pdf.

InternLM Team. Internlm3-8b-instruct. https://huggingface.co/internlm/internlm3-8b-instruct, 2025. Model card.

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

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay V. Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, Yuhuai Wu, Behnam Neyshabur, Guy Gur-Ari, and Vedant Misra. Solving quantitative reasoning problems with language models. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/18abbeef8cfe9203fdf9053c9c4fe191-Abstract-Conference.html.

Yiwei Li, Peiwen Yuan, Shaoxiong Feng, Boyuan Pan, Xinglin Wang, Bin Sun, Heda Wang, and Kan Li. Escape sky-high cost: Early-stopping self-consistency for multi-step reasoning. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=ndR8Ytrzhh.

M-A-P and ByteDance Seed. Supergpqa: Scaling LLM evaluation across 285 graduate disciplines. In Danielle Belgrave, Cheng Zhang, Laura N. Montoya, Hsuan-Tien Lin, Razvan Pascanu, Piotr Koniusz, Marzyeh Ghassemi, Nancy Chen, Iván Vladimir Meza Ruíz, and Arturo Loaiza-Bonilla (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025. URL http://papers.nips.cc/paper_files/paper/2025/hash/a3c5af1f56fc73eef1ba0f442739f5ca-Abstract-Datasets_and_Benchmarks_Track.html.

Math-AI. AMC 2023. Hugging Face Dataset, 2024. URL https://huggingface.co/datasets/math-ai/amc23. 40 problems from the 2023 American Mathematics Competitions.

Team Olmo, Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, Jacob Morrison, Jake Poznanski, et al. Olmo 3, 2026. URL https://arxiv.org/abs/2512.13961.

OpenAI. Openai o1 system card. https://openai.com/index/openai-o1-system-card/, 2024. Accessed: 2024-12-05.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026a. URL https://qwen.ai/blog?id=qwen3.5.

Qwen Team. Qwen3.6-35B-A3B: Agentic coding power, now open to all, April 2026b. URL https://qwen.ai/blog?id=qwen3.6-35b-a3b.

Georg Rasch. *Probabilistic Models for Some Intelligence and Attainment Tests*. Danish Institute for Educational Research, Copenhagen, Denmark, 1960.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. Gpqa: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*, 2024.

Charlie Victor Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM test-time compute optimally can be more effective than scaling parameters for reasoning. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=4FWAwZtd2n.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=1PL1NIMMrw.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Sihyeong Yeom, Geon Park, Geunyeong Jeong, Taewoong Yoon, Jaewook Lee, and Harksoo Kim. Is discrete difficulty sufficient? leveraging continuous difficulty for efficient self-consistency in llms. *CoRR*, abs/2608.24590, 2026. doi: 10.48550/ARXIV.2608.24590. URL https://doi.org/10.48550/arXiv.2608.24590.

Yifan Zhang and Team Math-AI. American invitational mathematics examination (aime) 2024, 2024.

Yifan Zhang and Math-AI Team. American invitational mathematics examination (aime) 2025, 2025. URL https://huggingface.co/datasets/math-ai/aime25.

A Experimental Setup and Data Processing

A.1 Generation and Trajectory Processing

For the core experiments and controlled constrained experiments, we generate 128 reasoning trajectories for each model–problem pair with temperature 0.7, top-$p = 0.8$, and a maximum output length of 4096 tokens. Generations terminated by the length limit are removed before analysis.

Final answers are extracted using a fixed rule-based procedure. Extracted answers are mapped to the answer-equivalence classes provided with each processed dataset, and answers assigned to the same class are aggregated. Trajectories for which no answer can be resolved are retained as incorrect responses for Rasch estimation and excluded from the construction of answer distributions.

For each case, the remaining answer clusters are ordered by empirical frequency. Geometry analyses use the resulting top-20-plus-tail representation, with zero padding when fewer than 20 clusters are observed.

B GEOMETRY ANALYSIS, NULL STRUCTURE, AND LAW ESTIMATION

B.1 PCA IMPLEMENTATION

The main geometric analysis is performed in a null-adjusted space. Matched null samples are first used to estimate a reference mean and covariance, which define the whitening transformation applied to the empirical ranked distributions. The transformed empirical vectors are then mean-centered, and their principal directions are obtained by singular-value decomposition.

The resulting PCA basis is fitted on the empirical distributions. Held-out null samples and constrained-answer distributions are transformed using the same frozen mapping and projected onto these empirical principal directions. This provides a common coordinate system for the geometric comparisons throughout the paper.

B.2 MATCHED NULL CONSTRUCTION AND NULL ADJUSTMENT

For each empirical case i, let K_i denote the number of observed answer clusters after filtering and N_i the number of trajectories contributing to its answer distribution. We generate the matched null as

$$q_i \sim \mathrm{Dirichlet}(\mathbf{1}_{K_i}), \qquad n_i \sim \mathrm{Multinomial}(N_i, q_i).$$

The Dirichlet concentration parameter is fixed to $\alpha = 1$. The sampled counts are ranked and converted to the same top-20-plus-tail representation as the empirical distributions.

Here, the observed K_i determines the dimension of the latent null distribution, while N_i preserves the finite-sampling scale of the corresponding empirical case. Multinomial sampling may leave some latent categories with zero counts, so the realized number of non-empty clusters may be smaller than K_i.

We generate 20 matched-null replicates for each case. The first 10 replicates are pooled to estimate the null mean and covariance used for whitening. The remaining 10 replicates provide the held-out null reference. Empirical distributions and held-out null samples are subsequently transformed by the same whitening map and evaluated in the same PCA space.

B.3 RASCH ESTIMATION AND CROSS-FITTING

Rasch observations are aggregated at the model–problem level. For each cell, all non-length-terminated trajectories contribute to the response count, with unresolved trajectories counted as incorrect. Let Y_{mi} and N_{mi} denote the number of correct and retained trajectories, respectively. We fit

$$Y_{mi} \sim \mathrm{Binomial}\left(N_{mi}, \sigma(\theta_m - \beta_i)\right).$$

The model abilities θ_m and problem difficulties β_i are estimated by alternating Newton updates of an L_2-regularized binomial log-likelihood, with regularization coefficient 1.0 for both parameter sets. Model abilities are initialized at zero, and item difficulties are initialized from a smoothed global response rate. After each iteration, the active model abilities are recentered to zero mean, with the same shift applied to the item parameters so that $\theta_m - \beta_i$ is preserved. Iteration continues until the maximum change in the fitted logits falls below 10^{-6}, with a maximum of 1,000 iterations.

The relative-capability coordinate is

$$r_{mi} = \theta_m - \beta_i.$$

We estimate these coordinates using deterministic five-fold cross-fitting over model–problem cells. Within each fold, the Rasch parameters are fitted on the other four folds and then used to assign r_{mi} to the held-out cells. This construction ensures that the correctness observations of a focal cell do not enter the fit used to obtain its own coordinate.

B.4 EMPIRICAL LAW ESTIMATION

The estimator of $F(r)$ follows the corresponding analysis setting.

For the full-data visualization, cases are ordered by r, summarized with fixed-case-count sliding windows, and represented using the leading five answer masses together with their remaining mass. The window centroids are smoothed with a ridge-regularized cubic spline.

The in-domain prediction experiment in Section 4 uses fixed-width 0.5 bins. Each occupied training bin is represented by the arithmetic mean of its answer distributions, and predictions between neighboring bins are obtained by linear interpolation. We evaluate this procedure using both top-5-plus-tail and top-20-plus-tail representations.

The transfer experiments in Section 5 use the full top-20-plus-tail representation together with the sliding-window spline estimator. For each held-out split, the fitted law and the mean-distribution baseline are computed from the training distributions.

C Additional Transition Analysis

C.1 Transition Estimation

Transition statistics are computed from fixed-case-count windows ordered by relative capability. Each window is located at the median r of its constituent cases.

The geometric response is defined as the speed of the centroid in the first three frozen null-adjusted principal components,

$$s_{\mathrm{shape}}(r) = \left\| \frac{d(\bar{z}_1, \bar{z}_2, \bar{z}_3)}{dr} \right\|_2 .$$

Correct-answer relocation is summarized from the derivatives of the probabilities assigned to the mutually exclusive correct-position states. Derivatives are evaluated at the observed window coordinates using numerical gradients and subsequently smoothed with a three-point kernel.

The component-wise changes in ranked probability mass provide the basis for the tail-to-leading-cluster and leading-cluster-to-top-1 patterns discussed in the main text.

C.2 Correct-Answer Rank

For each case, probability mass assigned to all correct-equivalent answer classes is combined before determining the rank of the correct answer. The correct rank is given by one plus the number of incorrect clusters with a strictly larger empirical count. Cases in which the correct answer is absent from the sampled responses form a separate state.

Correct-position probabilities are averaged over cases within each relative-capability window. Recoverable errors correspond to cases in which the correct answer has been sampled but another cluster occupies rank one.

D Generalization Details and Additional Results

D.1 Held-Out Evaluation Protocol

For leave-one-model-out evaluation, one model provides the test set and all remaining model–problem cases form the training set. Leave-one-dataset-out evaluation is constructed analogously by holding out one complete dataset. The training distributions determine both the fitted top-20-plus-tail law and the mean-distribution baseline.

These experiments use the casewise five-fold cross-fitted Rasch coordinates described in Appendix B.3. The identity-level split is applied to the fitting of $F(r)$: the held-out distributions are evaluated against a law estimated from the remaining distributions.

D.2 External Model–Benchmark Transfer

For HMMT25, the core law and baseline are fitted on the original development pool and fixed before external evaluation. Relative-capability coordinates for the external cases are obtained through an

anchored Rasch procedure. Abilities of core anchor models remain fixed while external item difficulties are estimated from their correctness responses. Abilities of new models are then estimated by five-fold item-level cross-fitting over the external benchmark.

The resulting coordinates place the external cases on the same relative-capability axis used by the frozen core law. The audited Qwen3.6-35B-A3B and DeepSeek external HMMT25 runs contain 64 trajectories per case, whereas the core experiments use 128.

E Constrained Answer-Space Experiments

E.1 Construction of Controlled Constrained Tasks

The controlled experiments use Qwen2.5-7B-Instruct and InternLM3-8B-Instruct on Minerva and OlympiadPhysics with $k \in \{3, 4, 5, 6, 7\}$.

For each model–problem pair, we rank the answer classes observed under open-answer sampling by empirical frequency and use the k most frequent classes as the candidate options for the corresponding k-choice variant. The original open-answer case is retained as the paired reference, so that the constrained and open distributions are compared for the same model and problem.

Each case uses a fixed candidate set and option ordering across all of its sampled trajectories. The prompt appends the candidate answers as multiple-choice options and asks the model to place the final option label in a boxed answer. Generation uses the same 128-sample configuration as the core experiments.

E.2 Projection and Geometric Comparison

The projection baseline applies top-k truncation and renormalization to the frozen open-answer law,

$$P_k(F(r)) = \frac{(F_1(r), \ldots, F_k(r))}{\sum_{j=1}^{k} F_j(r)}.$$

This provides the reference used to quantify how much of the constrained distribution can be explained by restricting the available answer space.

Geometric displacement is measured from matched constrained and open centroids in the first three dimensions of the frozen open-answer PCA basis. We additionally compute the corresponding Euclidean distance in the original 21-dimensional ranked-distribution representation as a sensitivity measure.

For GPQA-Diamond and SuperGPQA, the reference open distribution is obtained from the frozen core law evaluated at the calibrated relative-capability coordinate and projected to the corresponding answer-space size.