Same Benchmark, Same Subspace: Task-Selective Convergence in LLM Representations

JaeSeong Kim[1] **Suan Lee**[*1]

[1]Department of Computer Science, Semyung University, Jecheon-si, Republic of Korea,
{kjsqp1010,suanlee}@semyung.ac.kr

Abstract

Benchmark scores are the dominant lens through which progress in large language models is assessed, yet the implicit assumption that different benchmarks read out a shared internal structure has never been verified at the representation level. We introduce *readout subspace analysis*, a framework that extracts the low-dimensional subspace a benchmark uses to linearly discriminate correct answers from a model's frozen hidden representations, and compares these subspaces across models and tasks via a coordinate-invariant Gram cosine metric. Analyzing 50 models spanning six architecture families and varied fine-tuning strategies (SFT, DPO, RLHF, model merging) on three benchmarks (MMLU, MedQA, AGIEval), we find that **readout geometry is governed by the benchmark, not the model**. Within the same benchmark, readout subspaces are strongly aligned even across architecturally unrelated models (Gram cosine up to 0.87 versus a random baseline of $\sim 10^{-3}$), while subspaces across different benchmarks are nearly orthogonal. This convergence is graded: benchmarks sharing cognitive demands partially overlap in readout subspace, whereas those requiring qualitatively different reasoning occupy orthogonal regions. Furthermore, probe accuracy consistently meets or exceeds generation accuracy, indicating that benchmark scores reflect not only representational structure but also decoding efficiency. These findings recast benchmarks from passive measurement instruments to active structural constraints that selectively activate specific regions of representation space.

1 INTRODUCTION

Progress in large language models (LLMs) is predominantly narrated through benchmark scores. When a new model is released, its performance on standard benchmarks such as MMLU Hendrycks et al. [2021], MedQA Jin et al. [2021], and AGIEval Zhong et al. [2024] is reported, and improvements in these scores are interpreted as advances in model capability. A model that achieves uniformly high scores across multiple benchmarks is regarded as having acquired *general reasoning ability* beyond any single task. This score-based comparison has become the de facto evaluation framework across model leaderboards, technical reports, and academic publications.

For this interpretation to hold, a key assumption must be satisfied: that different benchmarks are reading out a shared internal structure within the model. That is, score improvements should reflect structural alignment in representation space or the sharing of underlying capability axes. Yet this assumption has never been directly verified at the level of representations. Whether the knowledge and reasoning demanded by each benchmark manifest as coherent geometric structures inside the model—and whether different models converge to similar structures for the same benchmark—remains unclear.

This work addresses these questions from the perspective of representation geometry. For every model–benchmark pair, we train a linear readout (probe) on frozen hidden representations and define the low-dimensional region spanned by the answer-discriminating directions as the *readout subspace*. By comparing these subspaces across 50 models spanning six architecture families, diverse pretraining corpora, and varied fine-tuning strategies, we ask:

> What determines the structure that reads out correct answers—the model or the benchmark?

Our key findings are as follows. **(1) Benchmark-driven convergence.** Within the same benchmark, readout subspaces

[*]Corresponding author

Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence (UAI 2026), PMLR 337:3101–3116.

are strongly aligned even across models with substantially different architectures and training trajectories (hundreds of times above the random baseline of $\sim 10^{-3}$ in a 4,096-dimensional space). **(2) Task-selective divergence.** Across different benchmarks, readout subspaces are nearly orthogonal, revealing the formation of task-specific structure independent of overall representational similarity. **(3) Graded structural sharing.** The relationship between benchmarks is not binary. MMLU and MedQA partially share subspace structure, whereas AGIEval exhibits minimal overlap with either. **(4) Latent knowledge beyond generation.** Probe accuracy meets or exceeds generation accuracy in every setting, suggesting that benchmark scores reflect a combination of representational structure and decoding efficiency.

These results prompt a reinterpretation of benchmark evaluation. Rather than passively measuring capability, benchmarks act as structural constraints that repeatedly activate specific subspaces in representation space. Consequently, aggregating benchmark scores is less akin to averaging a single ability and more akin to compressing distinct structural competencies onto a single axis.

Through readout subspace analysis, this work provides a framework for diagnosing inter-benchmark relationships not through score correlations, but at the level of representation geometry.

2 RELATED WORK

This work intersects with prior research on (i) representation analysis via linear probing, (ii) representational similarity measures, (iii) cross-model representational convergence, (iv) the gap between internal knowledge and generation, and (v) benchmark design and evaluation frameworks.

Representation analysis through probing. The probing paradigm—training an auxiliary classifier on frozen representations [Alain and Bengio, 2017]—has been extended to syntactic information diagnostics [Shi et al., 2016, Conneau et al., 2018], structural probe for syntax tree recovery [Hewitt and Manning, 2019], and edge probing [Tenney et al., 2019]. The debate over how to interpret probe accuracy—whether it reflects information intrinsic to the representations or is an artifact of probe capacity—has been addressed through control tasks [Hewitt and Liang, 2019], information-theoretic reformulations [Pimentel et al., 2020, Voita and Titov, 2020], and comprehensive surveys [Belinkov, 2022]. Our work departs from this literature by focusing not on probe accuracy per se, but on the geometric structure of the *readout subspace* defined by the probe weight matrix.

Representational similarity measures. Methods for comparing representations across models have progressed from SVCCA [Raghu et al., 2017] to PWCCA [Morcos et al., 2018] and CKA [Kornblith et al., 2019]. Principal angles [Björck and Golub, 1973] and Grassmann distances for measuring subspace relationships are also relevant.

Cross-model representational convergence. The phenomenon whereby representations from different models converge to shared structure has been documented through the Platonic Representation Hypothesis [Huh et al., 2024], SAE-based feature-level similarity [Bricken et al., 2023], and low-dimensional convergence in parameter space [Ainsworth et al., 2023]. The existence of universal directions is further supported by CCS [Burns et al., 2023] and the universality of truth subspaces [Bürger et al., 2024]. Whereas these studies address homogeneous convergence, our work reveals that *within* a converged representation space, distinct regions are selectively activated depending on the task. Readout subspaces converge strongly within the same benchmark, yet exhibit near-orthogonal separation across different benchmarks, newly exposing the existence of task-differentiated internal structure.

The gap between internal knowledge and generation. The discrepancy between the internal information of LLMs and their generated outputs has been reported through studies on self-calibration [Kadavath et al., 2022], self-critique [Saunders et al., 2022], inference-time intervention [Li et al., 2023], and the richness of internal states [Orgad et al., 2025]. The linear representation hypothesis [Park et al., 2024] further corroborates this view.

Benchmark design and evaluation frameworks. The benchmark lottery—where the choice of benchmarks determines conclusions [Dehghani et al., 2021]—perturbation sensitivity in MCQ evaluation [Alzahrani et al., 2024], and choices-only artifacts [Balepur et al., 2024] have all been documented. BIG-bench [Srivastava et al., 2023], HELM [Liang et al., 2023], and instance-level redundancy quantification [Wang et al., 2025, Zhang et al., 2025] are also related. While existing studies estimate inter-benchmark relationships through *correlations of model scores*, our work directly compares benchmarks at the *representation space level* via the geometric structure of readout subspaces, thereby circumventing the confound whereby score correlations conflate model ability with task difficulty.

3 METHOD

Our approach consists of three stages: (1) extracting the minimal structure that linearly reads out benchmark-specific knowledge from each model, (2) representing this structure in a form that is invariant to model-specific coordinate systems, and (3) verifying the existence of shared structure through cross-model and cross-benchmark comparisons. Each stage follows a single principle: we use *invariant representations* that do not depend on model-specific coordinate choices, ensuring that any observed convergence reflects

genuine structural alignment rather than coincidental similarity in coordinate systems.

3.1 LINEAR READOUT OF BENCHMARK-SPECIFIC KNOWLEDGE

For each model m and benchmark b, we train a linear classifier (probe) on the hidden representation $\mathbf{h}^{(\ell)}(x) \in \mathbb{R}^d$ at a designated layer ℓ. For a multiple-choice task with C_b classes, the readout weight matrix is defined as:

$$\mathbf{W}_{m,b}^{(\ell)} \in \mathbb{R}^{C_b \times d}. \tag{1}$$

Each row $\mathbf{w}_c$ represents the discriminative direction for the corresponding choice c, and only $\mathbf{W}$ is optimized while all model parameters remain frozen. Since a linear probe has no representational learning capacity of its own, the learned $\mathbf{W}$ reflects only information that is *already linearly encoded* in the model's representation space Alain and Bengio [2017]. A nonlinear probe, by contrast, risks learning structure absent from the model, thereby confounding model structure with probe structure Hewitt and Liang [2019]. Accordingly, $\mathbf{W}_{m,b}^{(\ell)}$ serves as a minimally sufficient description of "the directions that benchmark b uses to read out correct answers from layer ℓ of model m."

3.2 FROM READOUT WEIGHTS TO KNOWLEDGE SUBSPACES

We define the subspace spanned by $\mathbf{W}_{m,b}^{(\ell)}$ as the *readout subspace*—the region of representation space that a given benchmark demands from the model. Because different models operate in different coordinate systems, the absolute directions of the weight vectors are model-dependent. What we seek to compare is the *intrinsic geometry* of the readout—the relative relationships among discriminative directions. To this end, we use the Gram matrix:

$$\mathbf{G}_{m,b}^{(\ell)} = \mathbf{W}_{m,b}^{(\ell)\top} \mathbf{W}_{m,b}^{(\ell)}. \tag{2}$$

This Gram matrix possesses three key properties: (i) it is invariant to the ordering of class rows, making it unaffected by choice label assignment; (ii) it always resides in $\mathbb{R}^{d \times d}$ regardless of the number of classes C_b, enabling direct comparison between benchmarks with different answer cardinalities (e.g., MMLU with $C{=}4$ and AGIEval with $C{=}5$); and (iii) it fully preserves the geometric structure of the readout subspace.

3.3 SUBSPACE SIMILARITY VIA GRAM COSINE

The similarity between two readout subspaces is defined as the Frobenius cosine between their Gram matrices:

$$\mathrm{sim}(\mathbf{G}_1, \mathbf{G}_2) = \frac{\langle \mathbf{G}_1, \mathbf{G}_2 \rangle_F}{\|\mathbf{G}_1\|_F \, \|\mathbf{G}_2\|_F}. \tag{3}$$

Existing representational similarity metrics (CKA Kornblith et al. [2019], SVCCA Raghu et al. [2017]) measure similarity between model activations, answering the question "do models produce similar representations?" We instead ask: *"do benchmarks select similar regions of representation space?"* By comparing learned readout structures rather than activations, Gram cosine directly addresses this question.

3.4 ALIGNMENT MEASURES

For a benchmark pair (b_i, b_j), we define the cross-model average similarity as:

$$\mu^{(\ell)}(b_i, b_j) = \mathbb{E}_{m_i \neq m_j} \left[\mathrm{sim}\Big(\mathbf{G}_{m_i,b_i}^{(\ell)}, \, \mathbf{G}_{m_j,b_j}^{(\ell)} \Big) \right]. \tag{4}$$

When $b_i = b_j$, this yields the within-benchmark similarity; when $b_i \neq b_j$, it yields the across-benchmark similarity.

4 EXPERIMENTAL SETUP

Our experiments are designed to test a single core hypothesis: *What determines the structure that reads out correct answers—the model or the benchmark?* Rigorous verification of this hypothesis requires (1) that results reproduce across as diverse a set of models as possible, (2) that benchmarks with different cognitive demands are included so that we can distinguish task-specific convergence from global convergence, and (3) that analysis spans multiple network depths to trace when the relevant structure emerges.

4.1 MODELS: MAXIMIZING DIVERSITY

The greater the variation among models, the stronger the implication when convergence is observed. Following this principle, we select 50 LLMs along three axes of diversity: (i) **base architecture**—covering six or more distinct pretraining corpora, tokenizers, and architectures, including Mistral-7B, Llama-2-7B, Llama-3-8B, Qwen-7B/14B, Yi-6B, DeepSeek-Coder, OLMo, StableLM, and Baichuan2; (ii) **fine-tuning method**—encompassing fundamentally different optimization trajectories such as SFT, DPO, RLHF, and model merging; and (iii) **model scale**—ranging from 6B to 14B parameters.

These models are partitioned into two independent groups based on hidden dimension: **Group A** (38 models, $d{=}4096$, 32 layers) and **Group B** (12 models, $d{=}5120$, 40 layers). Since Gram cosine is defined only between models with identical d, all comparisons are conducted within groups, and we verify whether the patterns replicate independently across both. The full model list is provided in Appendix A.1.

4.2 BENCHMARKS: TRIANGULATING COGNITIVE DEMANDS

We select three multiple-choice benchmarks that systematically vary along domain breadth and cognitive demand type.

- **MMLU** (C=4, 57 subjects) Hendrycks et al. [2021]: An exam-based benchmark spanning a broad range of academic domains, requiring understanding of inter-concept relationships and contextual application. It exhibits high domain heterogeneity.
- **MedQA** (C=4) Jin et al. [2021]: Items drawn from the United States Medical Licensing Examination (USMLE), requiring clinical scenario interpretation and causal reasoning across disease–symptom–treatment relationships. It shares a knowledge-retrieval structure with MMLU but exhibits high within-domain cohesion.
- **AGIEval** (C=5) Zhong et al. [2024]: Composed of law and logic examination problems, centering on formal rule application and multi-step logical derivation. It demands a qualitatively different reasoning structure from knowledge-retrieval-oriented tasks.

Together, these three benchmarks independently vary domain diversity and reasoning type, providing the experimental contrasts needed to disentangle the cognitive origins of shared readout structure.

4.3 PROBE TRAINING AND PREPROCESSING

For each model m, benchmark b, and layer ℓ, we train a linear classifier on the hidden representation $\mathbf{h}^{(\ell)}(x) \in \mathbb{R}^d$. Hidden representations are extracted at the last token position of the input sequence, and only $\mathbf{W} \in \mathbb{R}^{C\times d}$ is optimized to minimize cross-entropy loss while all model parameters remain frozen. Training data follow the standard training split of each benchmark.

Prior to analysis, the readout weights are converted to effective weights that incorporate LayerNorm scaling: $\mathbf{W}_{\text{eff}} = \mathbf{W} \odot \boldsymbol{\gamma}^{(\ell)}$, where $\boldsymbol{\gamma}^{(\ell)}$ denotes the LayerNorm scale parameter of the corresponding layer. This ensures that the probe directions correctly reflect the original pre-normalization representation space rather than the normalized representation space.

4.4 LAYER SELECTION

For Group A, we analyze layers $\ell \in \{0, 5, 10, 15, 20, 25, 30, 32\}$; for Group B, $\ell \in \{0, 5, 10, 15, 20, 25, 30, 35, 40\}$. Layer 0 (immediately after embedding) serves as a baseline: high similarity at this stage is attributable to structural similarity in token embeddings and is not treated as evidence of benchmark-specific convergence. The five-layer interval enables tracking of when task-specific structure emerges and stabilizes across network depth, while the final layer (Group A: 32; Group B: 40) assesses the state of convergence in the model's terminal representation.

4.5 ANALYSIS PROTOCOL

We test the hypothesis through two complementary comparisons.

Within-benchmark vs. across-benchmark similarity. If $\mu^{(\ell)}(b, b)$ is significantly higher than $\mu^{(\ell)}(b_i, b_j)$ for $b_i \neq b_j$, this confirms that convergence is benchmark-specific rather than a byproduct of models being globally similar.

Cross-benchmark variation in convergence strength. We perform SVD on the collection of readout weights within each benchmark and compute the PC1 energy ratio $\rho = \sigma_1^2 / \sum_k \sigma_k^2$. A value of ρ substantially exceeding $1/N$ indicates that the readouts collapse into a low-dimensional subspace; differences in ρ across benchmarks reflect the internal cohesion of each benchmark.

Full experimental details are provided in Appendix A.

5 RESULTS

5.1 BENCHMARKS FORCE A SHARED READOUT STRUCTURE

We find that readout subspaces trained on the same benchmark converge regardless of the model. Figure 1 shows the layer-wise trajectory of Gram cosine, and Table 1 provides a quantitative summary. Table 1 reports values from the later layers. In the later layers of Group A (38 models), within-benchmark Gram cosine reaches MMLU 0.54, MedQA 0.78, and AGIEval 0.87, **exceeding the random expectation in high-dimensional space ($\sim 10^{-3}$) by several hundred fold**. These 38 models differ in base architecture (Mistral, Llama-2, Llama-3, Qwen, Yi, DeepSeek), training data, and fine-tuning method (SFT, DPO, RLHF, merge). The inter-model differences are extreme—for instance, a Mistral-7B-based SFT model and a Yi-6B-based DPO model share neither architecture, tokenizer, nor pre-training corpus. Yet for the same benchmark, they read out *the same directions* in representation space.

At the same time, across-benchmark similarity remains at the random level of $\sim 10^{-3}$ (MMLU↔AGIEval 0.001, MedQA↔AGIEval 0.001). This separation between within-benchmark and across-benchmark similarity indicates that convergence does not arise from global representational similarity across models, but rather constitutes a readout

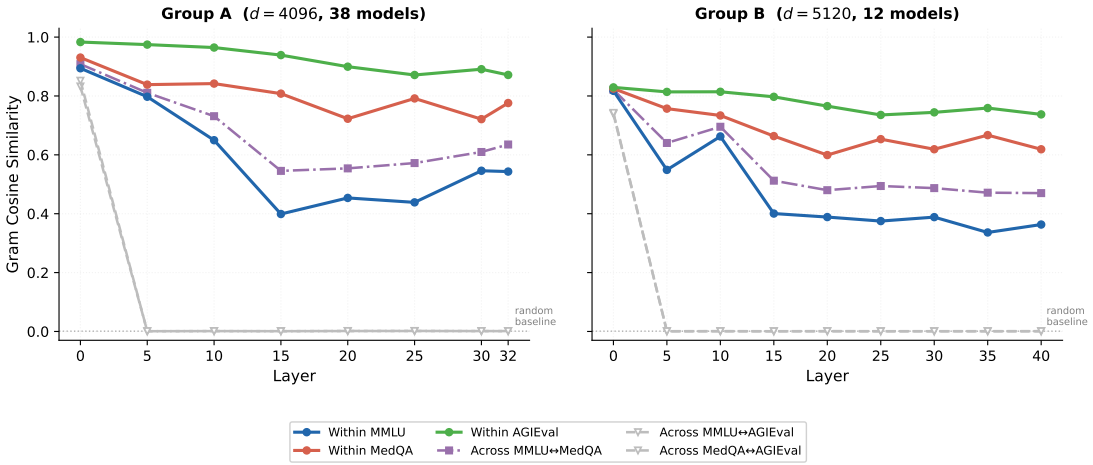

Figure 1: **Layer-by-layer Gram cosine trajectories.** Solid lines denote within-benchmark readout subspace similarity across models; dashed lines denote across-benchmark similarity. Within-benchmark similarity remains high after the initial layers, whereas across-benchmark similarity involving AGIEval drops to the random level ($\sim 10^{-3}$) at Layer 5. MMLU↔MedQA maintains structural overlap through the later layers, reflecting their shared reliance on factual knowledge retrieval. This pattern replicates independently in Group A (38 models) and Group B (12 models).

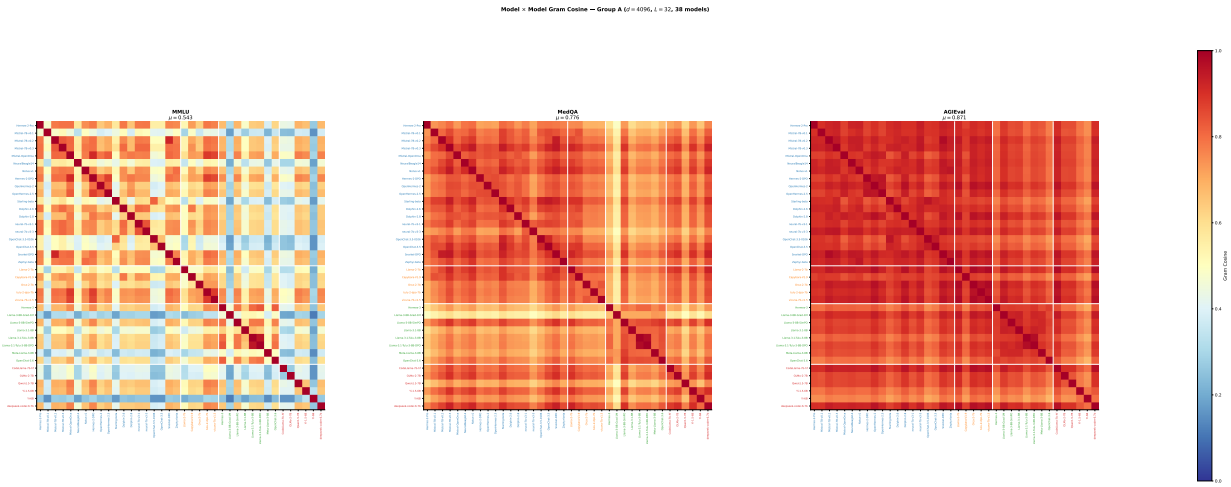

Figure 2: **Model×Model Gram cosine heatmaps — Group A (38 models, d=4096, last layer).** Each cell represents the Gram cosine between the readout subspaces of two models. Models are sorted by base architecture family, with white lines indicating family boundaries. (a) MMLU, the most heterogeneous benchmark, exhibits prominent within-family block structure with occasional low similarity (blue). (b) MedQA shows generally high similarity with some variation across model pairs. (c) AGIEval displays high similarity for nearly all model pairs, indicating the strongest convergence. μ denotes the off-diagonal mean.

structure specific to benchmark identity. That is, the same benchmark repeatedly activates similar subspaces across models with vastly different training histories, while different benchmarks select nearly orthogonal structures.

5.2 CONVERGENCE STRENGTH AS A FINGERPRINT OF BENCHMARK DESIGN

Not all benchmarks induce readout structure with equal strength. Convergence strength varies systematically across benchmarks (Figure 2). In the later layers of Group A, within-benchmark Gram cosine follows the ordering AGIEval 0.87 > MedQA 0.78 > MMLU 0.54. That is, the degree of cross-model alignment in readout subspaces trained on the same benchmark differs substantially by benchmark.

This pattern replicates independently in Group B (Figure 3): AGIEval 0.74 > MedQA 0.62 > MMLU 0.36. The same ordering recurs across two groups that differ in number of models, hidden dimension, and architectural composition.

Table 2: Layer-wise evolution of Gram cosine. At Layer 0, all pairs exhibit high similarity, but across-benchmark similarity involving AGIEval drops sharply from 0.83 to 0.001 at Layer 5. Benchmark-specific structure is then maintained throughout all subsequent layers. MMLU↔MedQA preserves structural sharing through the later layers.

	Group A (N=38, d=4096)						**Group B** (N=12, d=5120)					
	Within			Across			Within			Across		
L	MM	Md	AG	MM↔Md	MM↔AG	Md↔AG	MM	Md	AG	MM↔Md	MM↔AG	Md↔AG
0	.894	.931	.983	.908	.832	.852	.817	.825	.829	.821	.739	.743
5	.797	.839	.975	.811	.001	.001	.549	.757	.814	.640	.001	.001
10	.650	.842	.965	.731	.001	.001	.663	.734	.814	.695	.001	.001
15	.399	.808	.939	.545	.001	.001	.401	.664	.797	.512	.001	.001
20	.454	.723	.900	.554	.002	.001	.389	.599	.765	.480	.001	.001
25	.439	.792	.871	.572	.002	.002	.375	.653	.736	.494	.001	.001
30	.546	.721	.891	.610	.001	.001	.388	.619	.744	.487	.001	.001
last	.543	.776	.871	.635	.001	.001	.363	.619	.737	.470	.001	.001

last = Layer 32 (Group A) / Layer 40 (Group B). MM=MMLU, Md=MedQA, AG=AGIEval.

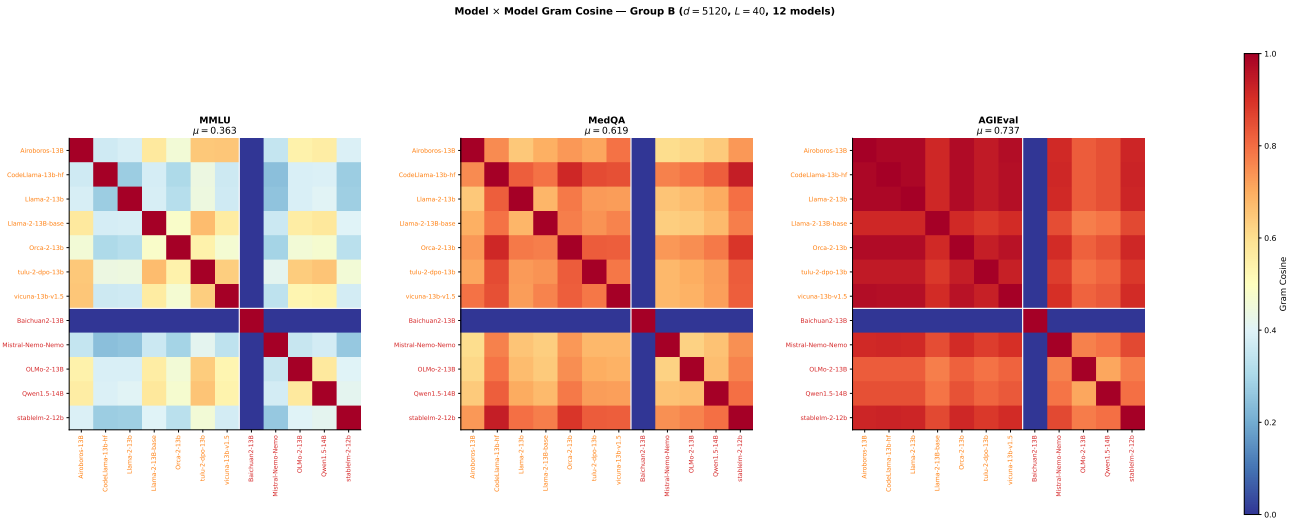

Figure 3: **Model×Model Gram cosine heatmaps — Group B (12 models, d=5120, last layer).** Replication in an independent group of 12 models separate from Group A. The cross-benchmark ordering of convergence strength (AGIEval > MedQA > MMLU) is exactly reproduced, with high cross-architecture similarity observed for AGIEval and MedQA.

These differences in convergence strength reflect variations in the cohesion of the reasoning structure each benchmark demands. Benchmarks with more uniform cognitive demands produce stronger cross-model alignment of readout subspaces, whereas more heterogeneous benchmarks exhibit relatively greater variance in readout structure across models.

5.3 BENCHMARKS SHARE SUBSPACES WHEN THEY SHARE COGNITIVE DEMANDS

The relationship between readout subspaces across benchmarks is not uniformly orthogonal but exhibits *selective sharing*. The across-benchmark Gram cosine for MMLU↔MedQA is 0.64, approximately 600 times the random expectation. Both benchmarks share factual knowledge retrieval as a core cognitive demand, and this shared demand manifests as overlapping readout subspaces in representation space.

In contrast, AGIEval is effectively orthogonal to both MMLU and MedQA, with similarity of $\sim 10^{-3}$. The logical reasoning structure demanded by AGIEval occupies an entirely different region of representation space from that used for knowledge retrieval.

MMLU spans 57 subjects, so different models may excel on different subsets, resulting in higher variance in readout structure. MedQA, by contrast, captures the

Table 1: Within-benchmark and across-benchmark Gram cosine in the later layers. Group A: 38 models, d=4096, layer 32. Group B: 12 models, d=5120, layer 40. Within-benchmark similarity exceeds the random expectation ($\sim 10^{-3}$) by several hundred fold, demonstrating that benchmarks induce model-invariant readout subspaces.

	Group A (N=38, L32)	**Group B** (N=12, L40)
Within-benchmark $\mu^{(\ell)}(b, b)$		
MMLU	$.543 \pm .174$	$.363 \pm .193$
MedQA	$.776 \pm .084$	$.619 \pm .285$
AGIEval	$.871 \pm .057$	$.737 \pm .335$
Across-benchmark $\mu^{(\ell)}(b_i, b_j)$		
MMLU $\leftrightarrow$ MedQA	$.635 \pm .152$	$.470 \pm .235$
MMLU $\leftrightarrow$ AGIEval	$.001 \pm .001$	$.001 \pm .000$
MedQA $\leftrightarrow$ AGIEval	$.001 \pm .001$	$.001 \pm .000$
Baselines		
Random subspace	$\sim 10^{-3}$	
Perm null (95pct)	.641	.488

knowledge-retrieval component present in MMLU *in purified form* within the single domain of medicine. Consequently, the MedQA readout aligns strongly with the knowledge-retrieval axis within MMLU, and this alignment is stronger than the inter-subject variance within MMLU itself. The fact that within-benchmark similarity for MedQA (0.78) exceeds that for MMLU (0.54) is consistent with this interpretation.

5.4 BENCHMARK-SPECIFIC STRUCTURE EMERGES EARLY AND PERSISTS

Table 2 shows the layer-wise evolution of readout structure. At Layer 0 (immediately after embedding), all benchmark pairs exhibit high similarity above 0.8. No inter-benchmark differentiation is observed at this stage; the readout structure is dominated by the shared token embedding basis.

Separation occurs at Layer 5. The across-benchmark similarity involving AGIEval (MMLU$\leftrightarrow$AGIEval, MedQA$\leftrightarrow$AGIEval) drops sharply from 0.83 to the 0.001 level. This indicates that differentiation of readout structure by reasoning type forms in the early layers of the network. This separation is then maintained throughout subsequent layers.

In contrast, MMLU$\leftrightarrow$MedQA sustains similarity in the range of 0.5–0.8 beyond Layer 5, showing that the structure shared by these two benchmarks is continuously preserved. The same pattern replicates in Group B.

5.5 MODELS KNOW MORE THAN THEY SHOW

Table 3 compares generation accuracy (Gen) and best-layer probe accuracy (Prb) for each model family. In every setting, probe accuracy matches or exceeds generation accuracy. This suggests that the information required for benchmark performance is linearly encoded in the model's hidden representations, while the generation process does not always fully exploit this information.

The gap is particularly pronounced for lower-performing models. For example, CodeLlama-7B records generation accuracy of .256 versus probe accuracy of .344 on MMLU, and Baichuan2-13B improves from .185 to .272 on AGIEval. In both cases, discriminative structure exists in the hidden representations but is not fully reflected in the generated output.

This observation suggests that benchmark scores can be interpreted as a combination of two factors: alignment with the readout subspace and the efficiency with which that structure is exploited during generation. Benchmark scores alone cannot disentangle these two contributions.

Table 3: Probe accuracy vs. generation accuracy, averaged across model families within each parameter group. N: number of models; K: number of distinct architecture families. Probe accuracy meets or exceeds generation accuracy in every setting, with larger gaps observed in lower-performing groups. This suggests that benchmark scores depend not only on the amount of knowledge retained but also on the efficiency of the decoding strategy.

			MMLU			MedQA			AGIEval		
Group	N	K	Gen	Prb	Δ	Gen	Prb	Δ	Gen	Prb	Δ
A (6–8B)	38	8	.485	.520	+.035	.420	.435	+.015	.386	.415	+.029
B (12–14B)	12	7	.410	.454	+.044	.353	.383	+.030	.316	.331	+.015
All	50	15	.467	.504	+.037	.404	.422	+.018	.369	.395	+.026

6 DISCUSSION

6.1 BENCHMARKS AS ACTIVE CONSTRAINTS, NOT PASSIVE MEASUREMENTS

This work has shown that benchmarks function not merely as collections of items, but as structural constraints that induce shared readout subspaces reproducibly across different models. The 50 models examined differ in architecture, pretraining data, and fine-tuning method. Yet readout structure converges within the same benchmark and is nearly orthogonal across different benchmarks.

This suggests that readout geometry is governed more strongly by benchmark identity than by model identity.

Prior work has documented global convergence of representations across models Huh et al. [2024]. Our findings

build on these observations by revealing that the interior of a converged representation space harbors task-differentiated structure. That is, independent of global representational similarity, task-specific readout subspaces exist.

6.2 WHAT BENCHMARK SCORES ACTUALLY MEASURE

The finding that probe accuracy exceeds generation accuracy calls for a reexamination of how benchmark scores are interpreted. The hidden representations linearly encode the information required for benchmark performance, yet the generation process does not always fully exploit this information. Benchmark scores can therefore be interpreted not as direct measurements of a single capability, but as a combination of two factors:

1. The degree to which the representation space is aligned with the readout subspace of the given benchmark.
2. The efficiency with which that structure is exploited during generation.

Scores alone cannot disentangle these two factors. A low score may stem not only from insufficient representational structure but also from inefficiency in the decoding strategy.

6.3 BENCHMARK COHERENCE AS MEASUREMENT PRECISION

The differences in convergence strength (AGIEval $>$ MedQA $>$ MMLU) are related to the cohesion of the structure each benchmark demands. Benchmarks with more uniform cognitive demands exhibit stronger cross-model readout alignment, whereas more heterogeneous benchmarks show greater variance in alignment.

Notably, the across-benchmark similarity for MMLU$\leftrightarrow$MedQA (0.64) exceeds the within-benchmark similarity for MMLU (0.54). This indicates that a single-domain benchmark can induce more cohesive structure than a mixed-domain benchmark. This result carries an implication for benchmark design: the more precisely defined the target structure, the more stable cross-model comparisons become.

6.4 TOWARD GEOMETRIC CHARACTERIZATION OF EVALUATION

Inter-benchmark relationships are typically analyzed through score correlations. However, score correlations cannot distinguish representational structure from difficulty effects. Gram cosine over readout subspaces directly compares the regions of representation space that benchmarks activate.

For instance, the finding that MMLU and MedQA share partially overlapping subspaces while AGIEval is nearly orthogonal to both reveals the redundancy and independence of evaluation axes at the structural level. This perspective opens the possibility of analyzing evaluation frameworks not through score-based comparisons but through the lens of representation geometry.

Acknowledgements

This work was supported by the Ministry of Science and ICT and the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (No. RS-2026-25498341).

References

Samuel K. Ainsworth, Jonathan Hayase, and Siddhartha Srinivasa. Git Re-Basin: Merging models modulo permutation symmetries. In *International Conference on Learning Representations (ICLR)*, 2023.

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. In *International Conference on Learning Representations (ICLR), Workshop Track*, 2017.

Norah Alzahrani, Hisham A. Alyahya, Yazeed Alnumay, Sultan Alrashed, Shaykhah Alsubaie, Yusef Almushaykeh, Faisal Mirza, Nouf Alotaibi, Nora Al-Twairesh, Areeb Alowisheq, M. Saiful Bari, and Haidar Khan. When benchmarks are targets: Revealing the sensitivity of large language model leaderboards. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024.

Nishant Balepur, Abhilasha Ravichander, and Rachel Rudinger. Artifacts or abduction: How do LLMs answer multiple-choice questions without the question? In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1): 207–219, 2022.

Åke Björck and Gene H. Golub. Numerical methods for computing angles between linear subspaces. *Mathematics of Computation*, 27(123):579–594, 1973.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Catherine Olsson, Tom Henighan, Danny Hernandez, Jared Kaplan, and Chris Olah. Towards monosemanticity: Decom-

posing language models with dictionary learning. *Transformer Circuits Thread*, 2023.

Lennart Bürger, Fred A. Hamprecht, and Boaz Nadler. Truth is universal: Robust detection of lies in LLMs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. In *International Conference on Learning Representations (ICLR)*, 2023.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2018.

Mostafa Dehghani, Yi Tay, Alexey A. Gritsenko, Zhe Zhao, Neil Houlsby, Fernando Diaz, Donald Metzler, and Oriol Vinyals. The benchmark lottery. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations (ICLR)*, 2021.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2019.

John Hewitt and Christopher D. Manning. A structural probe for finding syntax in word representations. In *Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2019.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The Platonic representation hypothesis. In *International Conference on Machine Learning (ICML)*. PMLR, 2024.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? A large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421, 2021.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Karina Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International Conference on Machine Learning (ICML)*. PMLR, 2019.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. Holistic evaluation of language models. *Transactions on Machine Learning Research (TMLR)*, 2023.

Ari S. Morcos, Maithra Raghu, and Samy Bengio. Insights on representational similarity in neural networks with canonical correlation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

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

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. In *International Conference on Machine Learning (ICML)*. PMLR, 2024.

Tiago Pimentel, Josef Valvoda, Rowan Hall Maudslay, Ran Zmigrod, Adina Williams, and Ryan Cotterell. Information-theoretic probing for linguistic structure. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020.

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. SVCCA: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

William Saunders, Catherine Yeh, Jeff Wu, Steven Bills, Long Ouyang, Jonathan Ward, and Jan Leike. Self-critiquing models for assisting human evaluators. *arXiv preprint arXiv:2206.05802*, 2022.

Xing Shi, Inkit Padhi, and Kevin Knight. Does string-based neural MT learn source syntax? In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2016.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research (TMLR)*, 2023.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R. Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R. Bowman, Dipanjan Das, and Ellie Pavlick. What do you learn from context? probing for sentence structure in contextualized word representations. In *International Conference on Learning Representations (ICLR)*, 2019.

Elena Voita and Ivan Titov. Information-theoretic probing with minimum description length. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020.

Shaobo Wang, Cong Wang, Wenjie Fu, Yue Min, Mingquan Feng, Isabel Guan, Xuming Hu, Conghui He, Cunxiang Wang, Kexin Yang, Xingzhang Ren, Fei Huang, Dayiheng Liu, and Linfeng Zhang. Rethinking LLM evaluation: Can we evaluate LLMs with 200x less data? *arXiv preprint arXiv:2510.10457*, 2025.

Zicheng Zhang, Xiangyu Zhao, Xinyu Fang, Chunyi Li, Xiaohong Liu, Xiongkuo Min, Haodong Duan, Kai Chen, and Guangtao Zhai. Redundancy principles for MLLMs benchmarks. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2025.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. AGIEval: A human-centric benchmark for evaluating foundation models. In *Findings of the Association for Computational Linguistics: NAACL*, 2024.

A DETAILED EXPERIMENTAL SETUP

A.1 MODEL SELECTION

Tables 4 and 5 present the full model lists for the two groups. The primary criterion for model selection is **diversity within the same hidden dimension**. A substantial number of models derived from the same base architecture are included (e.g., 19 models based on Mistral-7B); this is a deliberate design choice to ensure diversity in fine-tuning methods: it enables comparison of how fundamentally different optimization trajectories—SFT, DPO, RLHF, model merging—affect the readout subspace. At the same time, six or more distinct base architectures are included to verify that the observed convergence is not an artifact of a particular pretraining recipe. In particular, Qwen1.5-7B, Yi-6B, and DeepSeek-Coder in Group A differ from the Llama/Mistral families in pretraining data, tokenizer, and architectural details.

Table 4: Model list for Group A (d=4096, 32 layers, 38 models).

Family	Model	FT Method	Params
Mistral-7B	Mistral-7B-Instruct-v0.1	SFT	7.2B
	Mistral-7B-Instruct-v0.2	SFT	7.2B
	Mistral-7B-Instruct-v0.3	SFT	7.2B
	Hermes-2-Pro-Mistral-7B	SFT	7.2B
	OpenHermes-2-Mistral-7B	SFT	7.2B
	OpenHermes-2.5-Mistral-7B	SFT	7.2B
	Nous-Hermes-2-Mistral-7B-DPO	DPO	7.2B
	Mistral-7B-OpenOrca	SFT	7.2B
	NeuralBeagle14-7B	Merge	7.2B
	Notus-7B-v1	DPO	7.2B
	Starling-LM-7B-beta	RLHF	7.2B
	dolphin-2.6-mistral-7b-dpo	DPO	7.2B
	dolphin-2.8-mistral-7b-v02	DPO	7.2B
	neural-chat-7b-v3-1	SFT	7.2B
	neural-chat-7b-v3-3	SFT	7.2B
	openchat_3.5	C-RLFT	7.2B
	openchat-3.5-0106	C-RLFT	7.2B
	snorkel-mistral-pairrm-dpo	DPO	7.2B
	zephyr-7b-beta	DPO	7.2B
Llama-2-7B	Llama-2-7b-chat-hf	RLHF	6.7B
	Nous-Capybara-7B-V1.9	SFT	6.7B
	Orca-2-7b	SFT	6.7B
	tulu-2-dpo-7b	DPO	6.7B
	vicuna-7b-v1.5	SFT	6.7B
Llama-3-8B	Meta-Llama-3-8B-Instruct	SFT+RLHF	8.0B
	Llama-3.1-8B-Instruct	SFT+RLHF	8.0B
	Hermes-3-Llama-3.1-8B	SFT	8.0B
	Llama-3-Instruct-8B-SimPO	SimPO	8.0B
	Llama-3-8B-Instruct-Gradient-1048k	SFT	8.0B
	Llama-3.1-Tulu-3-8B	SFT	8.0B
	Llama-3.1-Tulu-3-8B-DPO	DPO	8.0B
	openchat-3.6-8b	C-RLFT	8.0B
Others	CodeLlama-7b-Instruct-hf	SFT	6.7B
	OLMo-2-1124-7B-Instruct	SFT+DPO	6.9B
	Qwen1.5-7B-Chat	SFT+RLHF	7.7B
	Yi-6B-Chat	SFT	6.1B
	Yi-1.5-6B-Chat	SFT	6.1B
	deepseek-coder-6.7b-instruct	SFT	6.7B

Table 5: Model list for Group B (d=5120, 40 layers, 12 models).

Family	Model	FT Method	Params
Llama-2-13B	Llama-2-13b-hf	— (base)	13.0B
	Llama-2-13b-chat-hf	RLHF	13.0B
	Airoboros-L2-13B-2.1	SFT	13.0B
	CodeLlama-13b-Instruct-hf	SFT	13.0B
	Orca-2-13b	SFT	13.0B
	tulu-2-dpo-13b	DPO	13.0B
	vicuna-13b-v1.5	SFT	13.0B
Others	Baichuan2-13B-Chat	RLHF	13.0B
	Mistral-Nemo-Instruct-2407	SFT	12.2B
	OLMo-2-1124-13B-Instruct	SFT+DPO	13.0B
	Qwen1.5-14B-Chat	SFT+RLHF	14.2B
	stablelm-2-12b-chat	SFT+DPO	12.1B

A.2 BENCHMARK DETAILS

Table 6 summarizes the composition and data splits of the three benchmarks.

Table 6: Benchmark composition and data splits.

	Classes	Domains	Few-shot	Total	Train	Val
MMLU	4	57 subjects	5-shot	14,042	11,233	2,809
MedQA	4	Medicine	0-shot	4,183	3,346	837
AGIEval	5	Law/Logic	0-shot	1,009	807	202

MMLU. We use the 57-subject multiple-choice questions from Hendrycks et al. Hendrycks et al. [2021]. For 5-shot evaluation, up to five examples from each subject's development set are used as the few-shot prefix. Few-shot examples are drawn from the same subject as the test item.

MedQA. We use the four-choice items based on the United States Medical Licensing Examination (USMLE) from Jin et al. Jin et al. [2021]. Evaluation is conducted in a 0-shot setting.

AGIEval. We use the law examination and logical reasoning items from Zhong et al. Zhong et al. [2024]. Items are in five-choice format and evaluated in a 0-shot setting.

Data splits. The full set of items for each benchmark is split into train and validation sets at an 80:20 ratio (seed=42). Only the train split is used for probe training, and the validation split is used for optimal epoch selection. No separate test set is used; validation accuracy is reported as the final probe performance. The rationale for this design is that the goal is not to optimize probe performance per se, but rather to analyze the *geometric structure of the learned weights*.

A.3 HIDDEN STATE EXTRACTION

For each model, hidden representations of the benchmark items are extracted as follows.

Hidden representation extraction. Hidden states are extracted from all transformer layers of the model, but only selected key layers are used for analysis:

- Group A (32 layers): $\ell \in \{0, 5, 10, 15, 20, 25, 30, 32\}$
- Group B (40 layers): $\ell \in \{0, 5, 10, 15, 20, 25, 30, 35, 40\}$

Here, $\ell = 0$ denotes the output immediately after token embedding (before the first transformer layer), and $\ell = L$ denotes the output of the final transformer layer.

At each layer, only the hidden state at the **last token position** is extracted to obtain $\mathbf{h}^{(\ell)}(x) \in \mathbb{R}^d$. This is the position used for next-token prediction in autoregressive models and represents where the model's final judgment on the item is concentrated.

A.4 PROBE ARCHITECTURE AND TRAINING

Architecture. The linear probe is a two-layer structure consisting of LayerNorm followed by a linear classifier:

$$f(\mathbf{h}) = \mathbf{W} \cdot \mathrm{LayerNorm}(\mathbf{h}) + \mathbf{b} \tag{5}$$

where $\mathrm{LayerNorm}$ has learnable scale ($\boldsymbol{\gamma}$) and shift ($\boldsymbol{\beta}$) parameters, and $\mathbf{W} \in \mathbb{R}^{C \times d}$, $\mathbf{b} \in \mathbb{R}^C$.

LayerNorm is included because hidden representations from different models may have different scales; LayerNorm normalizes these, improving the cross-model comparability of the classifier weights.

Training configuration. Table 7 summarizes the training hyperparameters.

Table 7: Probe training hyperparameters.

Parameter	Value
Optimizer	AdamW
Learning rate	5×10^{-4}
Weight decay	0 (default)
Epochs	3
Batch size	64
Loss	Cross-entropy
Epoch selection	Best validation accuracy

Validation accuracy is computed after each epoch, and the model weights from the epoch achieving the highest validation accuracy are selected as the final probe. The total number of probes trained is 50 models $\times$ 3 benchmarks $\times$ $\sim$8 layers $=$ $\sim$1,200.

A.5 EFFECTIVE WEIGHT COMPUTATION

The trained probe is a composite function of $\mathrm{LayerNorm} \to \mathrm{Linear}$. For analysis, we use the *effective weight* that reflects this composition.

The output of LayerNorm is (ignoring bias):

$$\mathrm{LayerNorm}(\mathbf{h})_i = \gamma_i \cdot \frac{h_i - \mu}{\sigma} \tag{6}$$

where γ_i is the learned scale parameter. The classifier output is:

$$\mathbf{W} \cdot \mathrm{LayerNorm}(\mathbf{h}) = \sum_i W_{c,i} \cdot \gamma_i \cdot \hat{h}_i = \mathbf{W}_{\text{eff}} \cdot \hat{\mathbf{h}} \tag{7}$$

where $\hat{\mathbf{h}}$ is the normalized hidden representation, and the effective weight is:

$$W_{\text{eff}}[c, i] = W[c, i] \cdot \gamma_i \tag{8}$$

That is, $\mathbf{W}_{\text{eff}} = \mathbf{W} \odot \boldsymbol{\gamma}$ (element-wise, broadcast along rows). This transformation yields the "actually operative directions" in which the LayerNorm scale has been absorbed into the classifier weights.

All Gram cosine analyses in this paper are performed on this effective weight $\mathbf{W}_{\text{eff}}$.

A.6 GENERATION ACCURACY COMPUTATION

For comparison with probe accuracy, generation accuracy for each model is measured as follows.

Decoding. Greedy decoding is used with a maximum generation length of 16 tokens. The first valid choice token (A, B, C, D, or E) appearing in the generated text is extracted as the answer.

Accuracy. Accuracy is computed by comparing the extracted answer against the ground-truth label. Evaluation is performed on the same validation split used for probes.

A.7 GRAM COSINE COMPUTATION DETAILS

Efficient computation. For $d = 4096$, directly constructing the Gram matrix $\mathbf{G} \in \mathbb{R}^{4096\times4096}$ and computing the Frobenius inner product requires $d^2 \approx 1.7 \times 10^7$ operations. Instead, we exploit the following identity:

$$\langle \mathbf{G}_1, \mathbf{G}_2 \rangle_F = \langle \mathbf{W}_1^\top \mathbf{W}_1,\ \mathbf{W}_2^\top \mathbf{W}_2 \rangle_F \tag{9}$$

$$= \mathrm{tr}\big(\mathbf{W}_1 \mathbf{W}_2^\top \mathbf{W}_2 \mathbf{W}_1^\top\big) \tag{10}$$

$$= \|\mathbf{W}_1 \mathbf{W}_2^\top\|_F^2 \tag{11}$$

Since $\mathbf{W}_1 \mathbf{W}_2^\top \in \mathbb{R}^{C_1\times C_2}$ and $C \leq 5$, this is a small matrix of at most $5 \times 5 = 25$ entries, requiring only $O(C_1 C_2 d)$ operations. The denominator is computed using the same trick: $\|\mathbf{G}\|_F = \|\mathbf{W}\mathbf{W}^\top\|_F$.

Number of pairs. For Group A (38 models), within-benchmark comparison yields $\binom{38}{2} = 703$ pairs, and across-benchmark comparison yields $38 \times 37 = 1{,}406$ pairs. For Group B (12 models), these are 66 and 132 pairs, respectively.

B PROMPT TEMPLATES

This appendix presents the exact format of the prompt templates used for the three benchmarks. The same template is applied to all models, and identical inputs are used for both generation accuracy (GenAcc) measurement and hidden state extraction.

G.1 MMLU (5-SHOT)

MMLU uses a 5-shot prompt per subject. Five examples are selected from the `dev` split of each subject to form the prefix.

MMLU Prompt Template

```
The following are multiple choice questions
(with answers) about {subject}.

{question_1}
A. {option_a}
B. {option_b}
C. {option_c}
D. {option_d}
Answer: {answer_1}

... (4 more few-shot examples) ...

{question}
A. {option_a}
B. {option_b}
C. {option_c}
D. {option_d}
Answer:
```

For evaluation, GenAcc is measured by checking whether the first generated token immediately after the `Answer:` token matches one of A/B/C/D. The last-token hidden state at the same position is used for probe training and evaluation.

G.2 MEDQA (0-SHOT)

MedQA is evaluated in a 0-shot setting without few-shot examples. The instruction `"Answer with the option letter only."` is included to elicit single-token generation.

MedQA Prompt Template

```
Question: {question}
Options:
A. {opa}
B. {opb}
C. {opc}
D. {opd}
Answer with the option letter only.
Answer:
```

G.3 AGIEVAL (0-SHOT)

AGIEval aggregates eight English configurations and uses a five-choice format (A–E). Some items include a passage; this field is omitted when no passage is present.

AGIEval Prompt Template

```
The following are multiple choice questions.
Choose the correct answer.

{passage}
Question: {question}
(A) {option_0}
(B) {option_1}
(C) {option_2}
(D) {option_3}
(E) {option_4}
Answer:
```