How Do LLM Evaluators Represent Human Competency? A Geometric and Causal Analysis

Anonymous ACL submission

Abstract

Human competency names the multi-dimensional pattern of traits (e.g., innovation, leadership, foresight) thought to underlie long-term professional excellence. Large language models (LLMs) are increasingly used as scalable judges here, but whether their scores reflect structured internal representations remains unclear. We systematically evaluate three open-source LLMs (ranging from 8B to 14B parameters) ranking performance on competency assessment when aligning with human expert judgments. To uncover the internal mechanisms of these LLM evaluators, we perform activation-wise linear probing to verify that competency traits are computed progressively and achieve peak performance in the deep layers of the models. Representational Similarity Analysis uncovers the topological structure of these traits, identifying critical core traits that function as structural "hubs". Steering based on these hub traits demonstrates superior modulation performance, outperforming the self steering baseline. Our work provides a geometrically grounded perspective on the latent representations of abstract human traits, enhancing the transparency and trustworthiness of LLM-as-a-judge systems.

1 Introduction

LLMs, characterized by their massive parameter scales and training over diverse corpora, have transcended mere statistical text completion to exhibit sophisticated reasoning, contextual understanding, and zero-shot generalization capabilities (Brown et al., 2020; Zoph et al., 2022). As LLMs increasingly demonstrate an ability to mirror human, they are increasingly deployed as automated evaluators across wide ranging fields, from academic essay scoring and code review to complex professional behavioral assessments (Gu et al., 2025).

Despite their growing utility as evaluators, the internal mechanisms by which LLMs produce judgments remain poorly understood, particularly in complex high-stakes assessments where evaluative criteria are inherently entangled and multidimensional. Moving beyond behavioral benchmarking, a central challenge is determining whether these outputs are grounded in structured internal representations rather than opaque latent heuristics. This lack of mechanistic interpretability substantially limits the trustworthiness of LLM-as-a-judge systems, as human experts cannot readily verify how high-dimensional competency signals are distilled into discrete evaluative scores (Gajcin et al., 2026; Rudin, 2019).

To systematically address this gap, this paper provides a comprehensive, layer-wise mechanistic look into the representations of LLM evaluators. Working with a robust dataset consisting of 583 evaluation samples mapped onto a detailed 15-dimension competency framework (Tab A1), we benchmark and dissect three representative open-source LLMs ranging from 8B to 14B parameters (Tab A2). Specifically, our investigation expands along the following five core research questions:

RQ1: Do LLM evaluators produce competency scores that align with human expert judgments?

RQ2: Are competency traits linearly decodable from the internal activations of LLM evaluators?

RQ3: What geometric structure do trait representations exhibit across network depth, and do traits form interpretable relational networks?

RQ4: Are these representational structures consistent across models with different architectures and training regimes?

RQ5: Can trait representations be causally manipulated via activation steering to shift evaluation outcomes?

2 Related Works

2.1 Human Competency Assessment

The assessment of high-level competency and professional potential has increasingly shifted from traditional cognitive screening toward multi-dimensional competency modeling. Rooted in McClelland's Competency Iceberg Model (McClelland, 1973), which serves as foundational paradigm of testing for behavioral outcomes rather than static intelligence, contemporary frameworks frequently isolate latent personal traits, motives, and core cognitive drivers. Spencer (Spencer and Spencer, 1993) subsequently operationalized these hidden traits by defining behavioral indicators across critical organizational domains, providing a structural blueprint for multi-criteria competency identification. To transform these qualitative competencies into a quantifiable psychometric instrument, empirical studies require a systematic measurement architecture that ensures both structural clarity and internal consistency (Evans, 1996). Consequently, the assessment framework developed in this study moves beyond singular evaluation metrics to establish an empirical measurement built upon four comprehensive dimensions: Innovation, Mission-Drive, Leadership, and Foresight.

2.2 LLM-as-a-Judge

The evaluation of Large Language Models (LLMs) has increasingly transitioned from surface-level lexical metrics like BLEU (Papineni et al., 2002) or ROUGE (Lin, 2004) toward advanced semantic evaluators. This paradigm was formalized by Zheng et al. (2023) through the LLM-as-a-Judge framework, demonstrating that frontier models like GPT-4 (OpenAI et al., 2024) could achieve human-level consistency and strong alignment with human preferences. Building upon this foundation, researchers have expanded these evaluation methodologies to a wider range of downstream tasks, including complex generation tasks (Liu et al., 2023), automated code review and correctness assessment (Tong and Zhang, 2024), and even high-stakes domains such as academic course evaluation (Chiang et al., 2024).

Despite their remarkable capabilities, LLM judges exhibit systematic vulnerabilities and cognitive shortcuts that heavily challenge their empirical reliability. Early literature laid the foundation by exposing superficial evaluator defects, such as positional preferences, verbosity heuristics, and self-enhancement tendencies (Wang et al., 2024). Recent studies reveal that LLM evaluators suffer from profound value discrepancies, ethnocentric alignment mismatches, and ideological variations when assessing latent human qualities (Chang et al., 2024; Chen et al., 2026; Rozado, 2024; Ye et al., 2024; Tao et al., 2024).

2.3 Interpreting and Steering Neural Representations

To decode knowledge representation in large language models, research has shifted from passive diagnostic analysis to active behavioral intervention. Diagnostically, linear probing (LP) and Sparse Autoencoders (SAEs) serve as the two parallel paradigms for feature discovery. While LP extracts explicit semantic directions via supervised classification (Alain and Bengio, 2018; Belinkov, 2021; Park et al., 2024), SAEs employ unsupervised dictionary learning (Cunningham et al., 2023; Bricken et al., 2023) to resolve representation superposition by disentangling dense activations into monosemantic features.

Transitioning to active manipulation, representation engineering leverages these discovered features as causal targets for weight-free activation steering during inference. Within this framework, Inference-Time Intervention (ITI) (Li et al., 2024) shifts internal activations along truthful directions to enhance veracity, while Activation Addition (Turner et al., 2024) scales behavioral control by injecting steering vectors derived from prompt differences. Together, the pipeline of LP/SAE discovery and subsequent activation steering enables rigorous causal verification of model behavior.

3 Methodology

3.1 Dataset Construction

The data construction is based on the Competency Framework (Table A1), a multi-tier structure comprising four primary dimensions and fifteen secondary dimensions, each accompanied by a precise operational definition. Secondary dimensions (or tasks) are denoted as A_{ij} or $Ability_i/Task_j$, where i represents the primary dimension and j denotes the secondary dimension. To construct the dataset, we gathered valid responses from 283 incoming doctoral students. To ensure a balanced label distribution, we augmented the dataset with 150 synthesized high-score samples and 150 low-score samples (generated by Gemini-3-pro and

Claude-sonnet-4.6), yielding a total of 583 samples. As detailed in Table A2, the empirical data (Original) exhibits a pronounced label imbalance, where responses are heavily clustered around specific intermediate score tiers while extreme scores remain drastically underrepresented. By incorporating the synthesized samples, the augmented dataset achieves a significantly more uniform and balanced distribution across all score levels, thereby preventing the model from developing classification biases.

3.2 Models

To ensure the robustness and generalizability of our proposed framework, we conduct a comprehensive evaluation across three representative open-source LLMs (Qwen2.5-14B-Instruct (Team et al., 2024), DeepSeek-R1-Distill-Qwen-14B (Guo et al., 2025) and Llama-3.1-8B (Grattafiori et al., 2024)) characterized by distinct parameter scales, architectural specializations, and linguistic proficiencies. By evaluating our approach across this diverse spectrum, we aim to demonstrate that the performance gains of our method are not artifacts of a specific model family but rather reflect a consistent improvement across varied LLM landscapes and linguistic contexts.

3.3 Latent Representation and Linear Probing

Modern LLMs primarily adopt a Transformer Decoder-only architecture, which processes sequences through a series of identical layers. Each layer consists of a Multi-Head Attention (MHA) mechanism and a Feed-Forward Network (FFN), linked together by residual connections.

$$x_{l+1} = x_l + \text{FFN}_l(\sum_{h=1}^{H} Q_l^h \text{Att}_l^h(P_l^h x_l)) \quad (1)$$

In (1), x_l, x_{l+1} represent the input and output vectors (hidden states) of the l-th Transformer layer. H is the total number of attention heads in the layer. $\text{Att}_l^h(\cdot)$ is the standard attention mechanism for the h-th head at layer l. P_l^h, Q_l^h are the projection matrices that map the hidden state into the head's subspace and $\text{FFN}_l(\cdot)$ represents the FFN at layer l.

For a given dataset of N prompts, we extract the activations from each attention head h at layer l, when generating the score token for specific task. Let $a_{l,i}^h$ represent the activation vector for the i-th prompt.

$$\min_{\theta,b} \sum_{i=1}^{N} \mathcal{L}\left(\sigma(\theta \cdot a_{l,i}^h + b), y_i\right) + \lambda\|\theta\|_2^2 \quad (2)$$

A linear classifier is trained on these frozen activations. The linear probing (2) consists of a weight vector θ and a bias b that minimizes a loss function. θ is the learned parameters that define the decision boundary. $\sigma(\cdot)$ is the softmax function mapping the output to a probability. $\lambda\|\theta\|_2^2$ is an L_2 regularization term used to prevent overfitting and ensure the probing generalizes well.

3.4 Representational Similarity Analysis

Representational Similarity Analysis(RSA) measures the consistency of class-level geometric structures across different task-specific feature spaces, ensuring that the resulting graph reflects the tasks' shared ability to cluster similar categories while isolating dissimilar ones.

For each task $t \in \{1, \ldots, T\}$, we first project the samples into a class-aware manifold. Let $\mathbf{X}^{(t)} \in \mathbf{R}^{N \times D}$ denote the feature matrix for task t, where N is the number of samples and D represents the hidden dimension. For each class $c \in \{1, \ldots, C\}$, the class prototype (centroid) $\mu_c^{(t)}$ is defined as the mean feature vector of all samples belonging to that class:

$$\mu_c^{(t)} = \frac{1}{|\mathcal{S}_c|} \sum_{i \in \mathcal{S}_c} \mathbf{x}_i^{(t)} \quad (3)$$

To capture the discriminative structure of task t, we construct a Representational Dissimilarity Matrix (RDM), denoted as $\mathbf{R}^{(t)} \in \mathbf{R}^{C \times C}$. Each entry $R_{ij}^{(t)}$ represents the geometric distance between the prototypes of class i and class j. Using the cosine distance to ensure scale invariance, we have:

$$R_{ij}^{(t)} = 1 - \frac{\mu_i^{(t)} \cdot \mu_j^{(t)}}{\|\mu_i^{(t)}\| \|\mu_j^{(t)}\|} \quad (4)$$

The matrix $\mathbf{R}^{(t)}$ serves as a signature of how task t perceives the relational boundaries between different semantic categories.

We define the edge weights of the task graph by computing the second-order similarity between the RDMs. For a pair of tasks u and v, the similarity W_{uv} is calculated as the Pearson correlation

between the upper triangular portions of their respective RDMs:

$$W_{uv} = \text{corr}\left(\text{vec}(\text{triu}(\mathbf{R}^{(u)})), \text{vec}(\text{triu}(\mathbf{R}^{(v)}))\right) \quad (5)$$

where $\text{triu}(\cdot)$ extracts the off-diagonal elements to avoid self-comparison redundancy, and $\text{vec}(\cdot)$ flattens the resulting values into a vector. The resulting adjacency matrix $\mathbf{W} \in \mathbf{R}^{T \times T}$ represents the Task-wise Relational Graph, where a high weight indicates that two tasks impose a similar topological structure on the sample labels.

3.5 Steering

Steering(6) is a technique that manipulates the model's latent representations toward specific subspaces during the forward pass. The intervention is applied to the selected attention heads as follows:

$$x_{l+1} = x_l + \sum_{h=1}^{H} Q_l^h \left(\text{Att}_l^h(P_l^h x_l) + \alpha \gamma_l^h \theta_l^h\right) \quad (6)$$

Respectively, α is the intervention strength that controls how much we shift the activations. γ_l^h is a layer-wise normalization factor calculated as the standard deviation of baseline activations to ensure numerical stability. $\theta_l^h$is the steering vector identified via linear probing for that specific head.

To comprehensively and systematically investigate the dynamics of representation steering, we define and evaluate two operational alignment modes based on the targeted dimensions of the interventions:

- **Self-steering Mode**: We intervene on a specific dimension t using its corresponding steering vector θ_t and evaluate the subsequent score changes on the identical dimension t.
- **Hub-steering Mode**: To evaluate cross-dimension transferability and systemic influence, we apply the intervention using the steering vector of a structural hub dimension θ_{hub} (dimension with highest graph centrality in Section 4.3), and observe the resultant score changes across neighboring dimensions $t' \neq$ hub.

4 Results

4.1 Alignment with Human Expert Ratings

To evaluate the alignment between LLM-based ratings and human annotations across diverse dimensions, we compute the statistical correlation between the model-assigned scores and the judgments of expert human raters. As shown in Table A2, the empirical results generally indicate a noticeable agreement between automated evaluations and human perspectives. 33 out of 45 model-task pairs achieved Kendall's W over 0.60, and 15 reached W over 0.70, indicating generally moderate-to-strong alignment between model's outputs and expert annotations. This prevalence of statistically verified correlations suggests that these open-source models possess a reasonable capacity to reflect human judgment when assessing specific behavioral and analytical traits. This alignment across a broad spectrum of dimensions indicates that leveraging advanced LLMs as evaluators can serve as a promising and scalable alternative to traditional human annotation, potentially mitigating annotation bottlenecks while maintaining acceptable evaluative fidelity (RQ1).

4.2 Localizing Traits Representations via Linear Probing

To investigate the capacity of LLMs to capture representations relevant to our tasks, we applied linear probing to the activations of individual attention heads across all network layers. The generated heatmaps Figure 1 and Figure A1 illustrate the linear probing accuracies for all 3 architectures. Due to space constraints, we limit our primary visual analysis to the Innovative Thinking task ($Ability_1/Task_0$). Heatmaps on both models reveal a consistent architectural phenomenon that deeper layers exhibit substantially higher linear probing accuracy on the relevant trait tasks compared to earlier layers. In the initial layers, the probing accuracy remains marginal, indicating poorly structured task features. As the residual stream progresses towards the final output layers, consecutive attention heads reach near-perfect accuracy. This monotonic increase strongly suggests that LLMs do not inherently possess static representations of competency traits at the token-embedding level. Instead, they gradually compute and refine these concrete semantic embeddings in a bottom-up manner as activations stack (RQ2).

To further validate and visually corroborate this progressive enhancement within the LLMs' embedding space, we applied t-SNE dimensionality reduction to cluster the sample-wise activations aggregated across all heads within specific target layers. As demonstrated in Figure 2 and Figure A2, the representation of latent traits in early layers is highly ambiguous. At this shallow stage, the

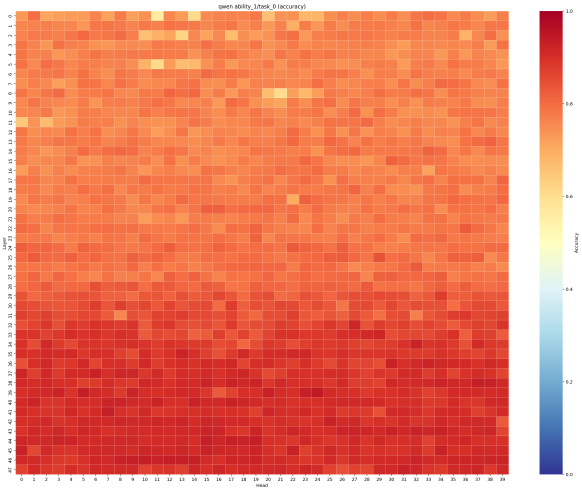


Figure 1: Heatmaps of Qwen's linear probing accuracy on activation from different heads for specific latent trait task (e.g. Innovative Thinking). For both models, X-axis is for heads from same layer and Y-axis is for layers.

latent space is heavily entangled, and only a minimal subset of attention heads exhibits rudimentary sensitivity capable of isolating samples with exceptionally high trait scores from the general distribution. However, as the depth increases, the representations mature and achieve significantly greater discriminative power. As shown in the intermediate layers, the clustering ability of the attention mechanisms improves drastically. While the decision boundaries for samples with intermediate scores remain somewhat entangled and indistinguishable, the network successfully creates well-defined, explicit boundaries that linearly separate the extremely high-scoring and low-scoring sample clusters. Ultimately, residing in the deep layers of the network, the model achieves a highly refined, fine-grained disentanglement of the trait continuum. At this terminal stage, the latent space effectively distinguishes samples across the entire scoring spectrum. Most notably, the representations of intermediary scores become progressively crystallized and the spatial overlap between adjacent score cohorts is significantly minimized, yielding distinct, clearly demarcated clusters.

4.3 Decoding Geometry of Traits Representations

While our linear probing analysis demonstrates that task-specific discriminative power monotonically increases with model depth, analyzing the inter-task relationship reveals a more intricate representational dynamic. To quantify this, we computed the mean pair-wise cosine similarity of activations among samples across all 15 selected tasks by layers and heads. As depicted in the heatmaps 3 and A3, the mean pair-wise cosine similarity across the 15 tasks does not follow a linear trend but instead exhibits a distinct U-shape trajectory as representations propagate. This contrasts with the monotonic rise in LP accuracy, suggesting that the model undergoes three distinct cognitive phases.

In the initial layers, the heatmap is saturated with high values. This extremely high similarity indicates that the early hidden states are heavily homogenized and undifferentiated. At this stage, the network is primarily processing low-level lexical, syntactic, and structural features of the input tokens, governed by model's and tokenizer's parameter prior. The contextual embeddings have not yet been strongly modulated by the specific task instructions. Consequently, the representations are globally entangled, leading to high cosine similarity across tasks, which intrinsically aligns with the poor linear probing accuracy observed at these shallow depths.

As computations progress into the intermediate layers, a significant transition occurs: the inter-task cosine similarity sharply declines. This U-shaped curve reflects a phase of functional specialization. During this stage, distinct attention heads diverge to extract and process heavily task-specific logic. The latent space bifurcates as different semantic pathways are engaged to tackle individual tasks independently. Because individual heads heavily encode isolated, non-overlapping task features, the pair-wise similarity across different task representations drops dramatically.

Remarkably, in the terminal layers, the similarity metrics rebound, returning to considerably high value. Although intermediate layers operate on highly specialized, independent pathways, the deeper layers re-integrate these dispersed features into a shared, cohesive semantic space. The model aggregates the specialized trait information into generalized, high-level cognitive abstractions just prior to the final unembedding projection. This subsequent increase in similarity indicates an enhancement in pair-wise relational capacity across tasks while the task features remain linearly separable for the classifier. As proven by the near-perfect terminal LP accuracy, their high-dimensional representations are spatially aligned within a unified, generic competency vector space capable of encod-

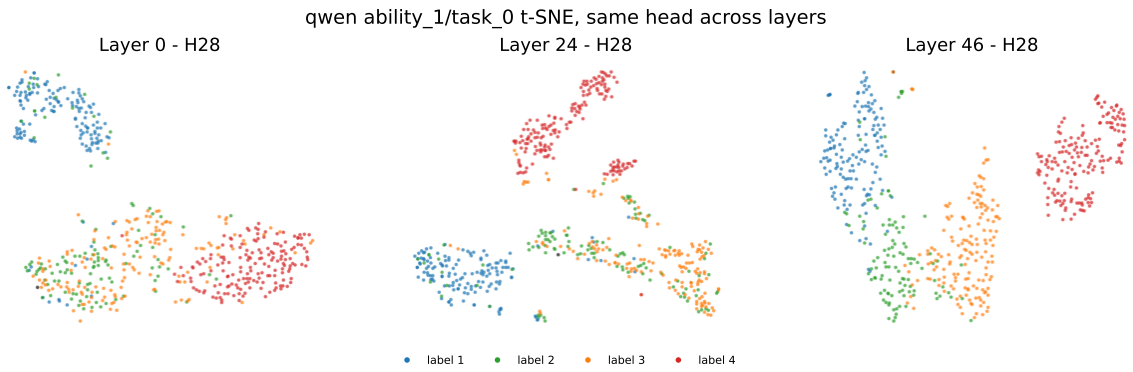


Figure 2: t-SNE clustering result of all 4-class samples on head 28 in layer 0/24/46, Qwen, Ability1/Task0(Innovative Thinking).

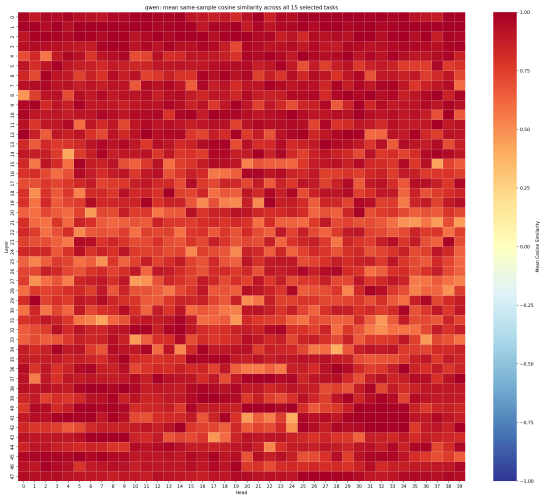


Figure 3: Heatmap of average cosine similarity among all 15 tasks on identical sample. X-axis denotes different heads from same layer and Y-axis denotes different layers.

ing complex semantic abstractions.

To systematically investigate the overlapping representational dynamics across different competence tasks, we analyzed the pairwise similarities of the model's latent embedding spaces. We computed the Pearson distance correlation across the 15 tasks, visualizing the resulting pairwise adjacency matrix as a heatmap and subsequently mapping it into a fully connected, weighted network graph to analyze its global topology.

As illustrated in the task distance correlation heatmap Figure 4 and Figure A4, the majority of the sub-tasks across varying ability categories exhibit exceptionally high inter-task correlations. This dense correlation block implies that the target LLM does not instantiate purely isolated circuits for each specific trait. Instead, it projects these evaluation tasks into a highly unified, shared latent subspace.

To elucidate the structural relationships among the evaluated capabilities, we model the correlation matrix as an undirected graph (Figure 4, Figure A5), where the edge weights correspond to the correlation. Node colors in the graph denote the betweenness centrality of each task, defined as a measure of a node's influence based on its connections to other nodes. Specifically, the task Ethical Commitment ($Ability_2/Task_2$) exhibits maximum centrality, acting as structural hubs. This task likely represents the most prototypical cognitive abilities that share representational pathways with the rest of the tasks (RQ3, RQ4).

4.4 Manipulating Traits Scoring via Activation Steering

To further elucidate the representational associations between hub task with others, we adopt a steering paradigm to manipulate the scoring process. Table 1 illustrates the impact of representation steering across diverse dimensions and configurations. Our results demonstrate that steering manipulation consistently modulates model outputs in a directionally predictable manner. Across all 15 latent traits, applying a positive steering vector leads to a marked increase in target attribute logits, while negative steering results in a corresponding suppression. This bidirectional control confirms that our extracted steering vectors effectively capture the semantic polarity of the targeted attributes in the representation space.

Crucially, the effectiveness of hub-steering aligns with the theoretical framework established in our related work. The data reveal that the identified hub nodes serve as robust control points for

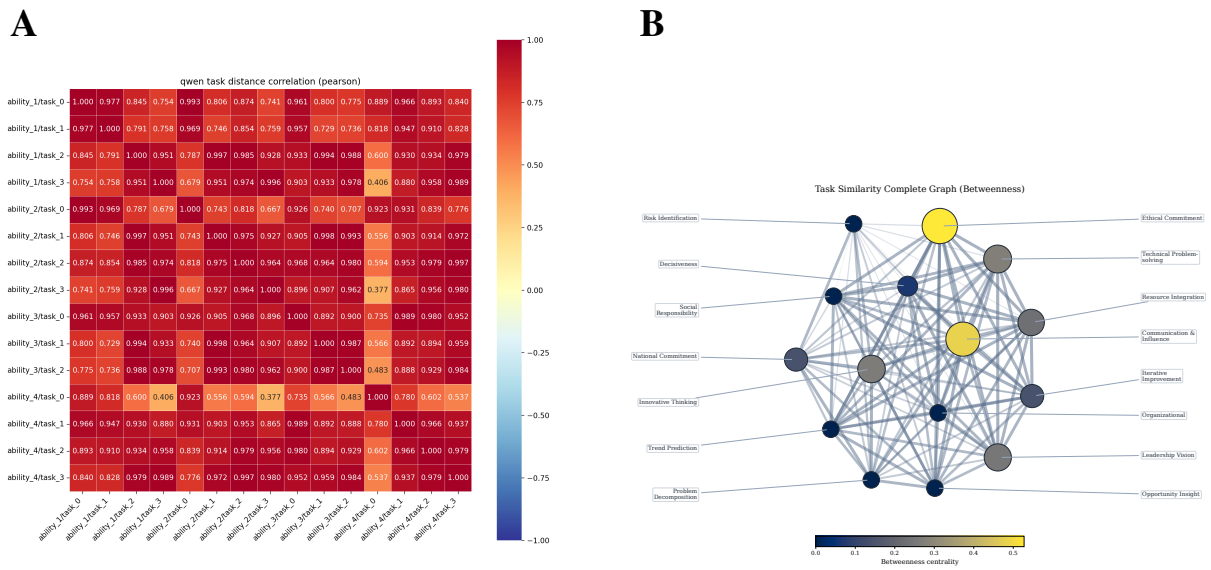


Figure 4: (A) A heatmap illustrating the RDM results using Pearson correlation matrix of task distances for the Qwen model. (B) A network graph visualizing the relationships between all 15 tasks. Each node represents a specific capability while connecting edges denoting their structural similarities. The fill color of each node indicates its betweenness centrality.

broader behavioral traits. This finding empirically validates the inter-dimensional relationships found in traditional leadership analysis, where foundational ethical orientations act as a primary driver for organizational and innovative outcomes.

Furthermore, we observe that, across the three analyzed models, hub-steering often yields a larger modulation magnitude than self-steering on paired task comparisons (Wilcoxon signed-rank test on task-averaged aligned effects: two-sided p = 0.0129; one-sided p = 0.0064; Figure 5). This suggests that hub nodes may capture more global semantic axes with stronger downstream influence over subordinate traits than the traits' own localized representations. Notably, this effect is especially pronounced in DeepSeek-Qwen (two-sided p = 6.1e-05), while Qwen and Llama-3.1-8B show the same qualitative trend in some tasks but do not individually reach significance. Taken together, these results support hub-based organization as a plausible cross-model tendency, rather than a uniformly confirmed effect in every architecture.

Furthermore, we observe that in several specific tasks, hub-steering achieves a greater magnitude of modulation than self-steering (Figure 5). This suggests that hub nodes may represent more fundamental, high-level semantic axes that possess stronger latent influence over subordinate traits than the traits' own localized representations. These steering effects remain consistent across all architectures, indicating that the hub-based structural organization of these latent traits is a robust, model-agnostic phenomenon within large language models (RQ5).

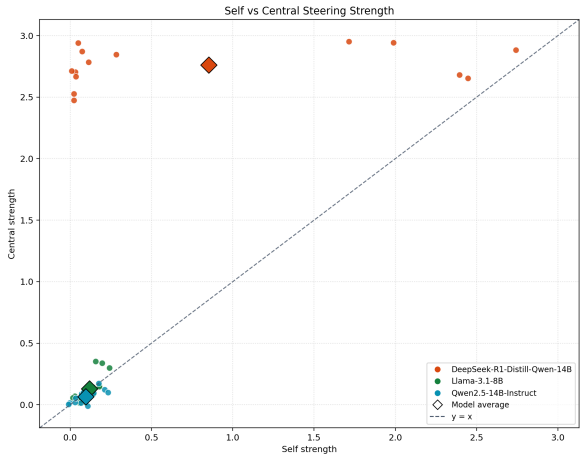


Figure 5: Comparison of steering strengths between self-steering and hub-steering across all tasks and models. Each point represents a task-specific steering result, where the x-axis denotes the self-steering strength and the y-axis denotes the hub-steering strength.

5 Conclusion

In this study, we provided a mechanistic characterization of the latent geometry governing competency trait representations within Large Language Models. Through a combination of linear probing and representational similarity analysis, we demonstrated that LLMs do not rely on static embeddings

Task	Qwen	DeepSeek	Llama
Self (H=16, $\|A\|$=100)			
Innovative Thinking	+0.075 / -0.071	+1.149 / -2.828	+0.242 / -0.072
Problem Decomposition	+0.194 / -0.095	+2.245 / -2.648	+0.250 / -0.033
Technical Problem-solving	+0.212 / -0.213	+0.000 / -0.062	+0.218 / -0.061
Iterative Improvement	-0.013 / -0.006	+0.146 / -1.541	+0.159 / -0.037
National Commitment	+0.115 / -0.024	+0.000 / -0.098	+0.022 / -0.094
Leadership Vision	+0.068 / -0.066	+0.075 / -0.072	+0.037 / -0.021
Ethical Commitment	+0.077 / -0.088	+0.046 / +0.000	+0.024 / -0.097
Social Responsibility	+0.060 / +0.000	+0.143 / -0.083	+0.094 / -0.092
Communication & Influence	+0.194 / -0.273	+2.600 / -2.882	+0.125 / -0.125
Organizational	+0.088 / -0.128	+0.000 / -0.071	-0.006 / -0.040
Decisiveness	-0.018 / +0.000	+0.568 / +0.000	+0.083 / -0.047
Risk Identification	+0.233 / -0.118	+0.429 / -3.000	+0.240 / -0.244
Opportunity Insight	+0.155 / -0.083	+0.020 / +0.000	+0.329 / -0.064
Trend Prediction	+0.136 / -0.055	+2.299 / -2.491	+0.321 / -0.037
Resource Integration	+0.056 / -0.012	+0.022 / -0.025	+0.244 / -0.115
Central (H=16, $\|A\|$=100)			
Innovative Thinking	+0.116 / -0.036	+2.885 / -3.000	+0.591 / -0.113
Problem Decomposition	+0.129 / -0.059	NA / NA	+0.077 / -0.094
Technical Problem-solving	+0.113 / -0.134	+2.781 / -2.625	+0.099 / -0.189
Iterative Improvement	+0.019 / -0.012	+2.888 / -2.639	+0.083 / -0.062
National Commitment	+0.154 / -0.024	+2.881 / -3.000	-0.032 / -0.062
Leadership Vision	-0.014 / -0.042	+2.849 / -2.895	+0.055 / -0.086
Ethical Commitment	NA / NA	+2.517 / -2.536	NA / NA
Social Responsibility	+0.040 / +0.000	+3.200 / -2.369	+0.029 / -0.011
Communication & Influence	+0.101 / -0.099	+2.764 / -3.000	+0.109 / -0.143
Organizational	-0.088 / -0.067	+2.336 / -3.000	+0.006 / -0.107
Decisiveness	-0.006 / -0.012	+2.694 / -3.000	+0.032 / -0.037
Risk Identification	+0.189 / -0.157	+2.905 / -3.000	+0.320 / -0.280
Opportunity Insight	+0.128 / -0.045	+2.640 / -2.786	+0.589 / -0.088
Trend Prediction	+0.054 / -0.037	+2.612 / -2.748	+0.263 / -0.053
Resource Integration	+0.099 / -0.006	+2.778 / -2.169	+0.252 / -0.044

Table 1: Steering results for 3 models across 15 tasks with number of heads set to 16 and α set to ± 100, where NA represents hub traits not counted in hub-mode.

but instead gradually compute abstract human qualities through a bottom-up refinement process. This journey progresses from early-layer homogenization to functional specialization, ultimately converging into a shared, re-integrated latent representations where disparate traits are spatially aligned.

Our network-based analysis identified critical structural hub nodes. Most prominently foundational dimensions like Ethical Commitment exhibit high centrality and serve as robust control points for broader behavioral traits. Empirically, activation steering through these hub nodes consistently performs as well as localized self-steering in modulating model outputs, effectively validating the inter-dimensional dependencies long theorized in human behavioral research. The robustness of these findings across diverse architectures suggests that such geometric organization is a model-agnostic feature of LLMs. By transforming black-box evaluations into an interpretable and geometrically grounded framework, our work offers a pathway toward mechanistic transparency in LLM's scoring process

Limitations

Our study is primarily focused on dense, decoder-only models of moderate parameter scales (8B-14B). Future research is required to determine whether these geometric patterns persist in significantly larger model architectures or sparse models (e.g. MOE models), where representation sparsity may alter the latent topology.

Our methodological framework relies on labels generated by LLMs to train linear probes and validate steering effects. While the "LLM-as-a-Judge" paradigm is increasingly standard, it is susceptible to inherent systematic biases and prompt sensitivity. Although our internal probing aims to bypass surface-level output biases, the initial "ground truth" remains a latent construct of the scorer models rather than an absolute psychometric measurement from human experts.

The empirical foundation of this study is subject to linguistic and demographic constraints: the dataset was collected using Chinese-language questionnaires from a cohort of Chinese students. Since competency traits are inherently socialcultural constructs, their latent representations may vary across different languages and cultural frameworks

To address data imbalance, we incorporated synthetic high- and low-score samples generated by LLMs. While this augmentation facilitates robust probing, it may introduce a "model-in-the-loop" bias, where the identified latent geometry partially reflects the model's internal generative logic and structural artifacts rather than a pure mirror of human psychology.

Acknowledgments

LLM tools were utilized for the writing, coding and visualization components of this paper.

References

Guillaume Alain and Yoshua Bengio. 2018. Understanding intermediate layers using linear classifier probes. *Preprint*, arXiv:1610.01644.

Yonatan Belinkov. 2021. Probing classifiers: Promises, shortcomings, and advances. *Preprint*, arXiv:2102.12452.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, and 1 others. 2023. Towards monosemanticity: Decomposing language models with dictionary learning. Technical report, Anthropic.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. 2024. A survey on evaluation of large language models. *ACM Trans. Intell. Syst. Technol.*, 15(3).

Jieying Chen, Karen de Jong, Andreas Poole, Jan Burakowski, Elena Elderson Nosti, Joep Windt, and Chendi Wang. 2026. Uncovering political bias in large language models using parliamentary voting records. *Preprint*, arXiv:2601.08785.

Cheng-Han Chiang, Wei-Chih Chen, Chun-Yi Kuan, Chienchou Yang, and Hung-yi Lee. 2024. Large language model as an assignment evaluator: Insights, feedback, and challenges in a 1000+ student course. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2489–2513, Miami, Florida, USA. Association for Computational Linguistics.

Hoagy Cunningham, Aidan Eisma, Logan Chan, Tilly Turnage, and Lee Sharkey. 2023. Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600*.

J. D. Evans. 1996. *Straightforward statistics for the behavioral sciences*. Thomson Brooks/Cole Publishing.

Jasmina Gajcin, Erik Miehling, Rahul Nair, Elizabeth M. Daly, Radu Marinescu, and Seshu Tirupathi. 2026. Interpreting LLM-as-a-judge policies via verifiable global explanations.

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

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. 2025. A survey on llm-as-a-judge. *Preprint*, arXiv:2411.15594.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang,

Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, and 175 others. 2025. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645:633–638.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2024. Inference-time intervention: Eliciting truthful answers from a language model. *Preprint*, arXiv:2306.03341.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

David C. McClelland. 1973. Testing for competence rather than for "intelligence". *American Psychologist*, 28(1):1–14.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Kiho Park, Yo Joong Choe, and Victor Veitch. 2024. The linear representation hypothesis and the geometry of large language models. *Preprint*, arXiv:2311.03658.

David Rozado. 2024. The political preferences of llms. *Preprint*, arXiv:2402.01789.

Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215.

Lyle M. Spencer and Signe M. Spencer. 1993. *Competence at work: Models for superior performance*. John Wiley & Sons.

Yan Tao, Olga Viberg, Ryan S Baker, and René F Kizilcec. 2024. Cultural bias and cultural alignment of large language models. *PNAS Nexus*, 3(9).

Qwen Team, An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, and 1 others. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*. Qwen2.5-14B-Instruct.

Weixi Tong and Tianyi Zhang. 2024. Codejudge: Evaluating code generation with large language models. *Preprint*, arXiv:2410.02184.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. 2024. Steering language models with activation engineering. *Preprint*, arXiv:2308.10248.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. 2024. Large language models are not fair evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9440–9450, Bangkok, Thailand. Association for Computational Linguistics.

Jiayi Ye, Yanbo Wang, Yue Huang, Dongping Chen, Qihui Zhang, Nuno Moniz, Tian Gao, Werner Geyer, Chao Huang, Pin-Yu Chen, Nitesh V Chawla, and Xiangliang Zhang. 2024. Justice or prejudice? quantifying biases in llm-as-a-judge. *arXiv preprint arXiv:2410.02736*.

Lianmin Zheng, Wei-Lin Chiang, Ion Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhen Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pages 46595–46623.

Barret Zoph, Colin Raffel, Dale Schuurmans, Dani Yogatama, Denny Zhou, Don Metzler, Ed H. Chi, Jason Wei, Jeff Dean, Liam B. Fedus, Maarten Paul Bosma, Oriol Vinyals, Percy Liang, Sebastian Borgeaud, Tatsunori B. Hashimoto, and Yi Tay. 2022. Emergent abilities of large language models. *TMLR*.

A Competence Framework

To provide a granular and structured baseline for competence evaluation, this study establishes a comprehensive competence framework comprising four primary dimensions and fifteen secondary dimensions. Table A1 presents the detailed definitions and operational conceptualizations of these dimensions, which collectively map the essential attributes required for competence evaluation.

B Data Distribution

To validate the robustness of the evaluation framework and mitigate potential data biases, we analyze the score distributions across various Large Language Models (LLMs) using both the original and augmented datasets. Table A2 delineates the empirical sample counts aggregated across distinct score tiers (Scores 0–1, 2, 3, and 4) for Qwen, Llama, and DeepSeek.

C Kendall's W between models and human experts

To rigorously quantify the alignment between LLMs evaluations and high-level human expertise, we compute the Kendall's W coefficient of concordance across the 15 competency dimensions. Table A3 reports the statistical agreement between human expert panels and three leading LLMs: DeepSeek, LLaMA, and Qwen, where the asterisk annotations indicates the corresponding statistical significance tiers

D Complementary Linear Probing Accuracy

Complementary figures in Section 4.2 are shown as follows (Figure A1), due to page limit.

E t-SNE of Model Representation

To investigate the inner mechanisms of Large Language Models in encoding competence-related features, we extract the hidden states of Qwen at different transformer layers and project them into a low-dimensional space using t-Distributed Stochastic Neighbor Embedding (t-SNE). Figureure A2 illustrates the representation trajectories of the 4-class proficiency samples across Layer 0 (input/embedding level), Layer 24 (intermediate level), and Layer 46 (deep level) specifically for $Ability_1/Task_0$ (*Innovative Thinking*).

F Complementary Figureures in Representation Similarity Analysis

Complementary figures in Section 4.3 are shown as follows (Figure A3 and Figure A4), due to page limit.

G Comprehensive Graph Visualization

Comprehensive graph in Section 4.3 are shown as follows (Figure A5).

H Ablation

To evaluate the operational robustness and parametric sensitivity of our activation steering methodology, we conduct a rigorous ablation study on Qwen model across 15 sub-tasks. We primarily ablate two pivotal hyperparameters: the number of intervention heads ($H \in \{16, 32\}$) and the steering scale factor ($|A| \in \{50, 100\}$), while benchmarking two distinct architectural configurations: Self-steering (Self) and Hub-steering (Central). Table A4 reports the dual-directional score shifts, where the prefix values ($+\alpha$) denote the competency alignment effect when steering low-score samples positively, and the suffix values ($-\alpha$) represent the degradation effect when steering high-score samples negatively. We choose setting $(head = 16, \alpha = 100)$ as the final setting for its best overall performance.

Primary Dimension	Secondary Dimension	Description
Innovation	Innovative Thinking	Capable of transcending conventional paradigms to propose disruptive technical concepts or application scenarios, thereby driving original innovation within the field.
	Problem Decomposition	Systematically applies first-principles thinking to deconstruct multi-faceted challenges into irreducible, actionable units.
	Technical Problem-Solving	Demonstrates the persistence and leadership required to mobilize teams in overcoming critical technical bottlenecks through intensive research and rigorous experimentation.
	Iterative Improvement	Embraces and implements rapid prototyping and feedback loops to facilitate continuous refinement, ultimately achieving substantial advancements in key performance indicators.
Mission Drive	National Commitment	Aligns individual and collective aspirations with national strategic development, striving to enhance national core competitiveness through the advancement of Artificial Intelligence.
	Leadership Vision	Possesses the strategic foresight to shape industry direction and the charismatic leadership required to unify teams toward a common, ambitious vision.
	Ethical Commitment	Strictly adheres to research integrity and academic norms, proactively identifying and mitigating ethical risks throughout the entire lifecycle of technology development.
	Social Responsibility	Driven by a sense of mission to address grand societal challenges through AI innovation, ensuring that technological outputs deliver inclusive benefits and positive social impact.
Leadership	Communication & Influence	Effectively articulates complex research findings to heterogeneous audiences with clarity and persuasion; leverages professional charisma, logical reasoning, and emotional resonance to inspire alignment with collective objectives.
	Organizational	Adept at designing and optimizing high-performance organizational structures; excels in resource allocation and task delegation, emerging as a strategic nucleus that drives research teams toward mission success.
	Decisiveness	Demonstrates the capacity to synthesize available information into decisive strategic actions, balancing immediate exigencies with long-term developmental goals.
Foresight	Risk Identification	Proactively anticipates and evaluates multi-dimensional risks across technical and market landscapes; excels in developing robust contingency plans to ensure project resilience.
	Opportunity Insight	Discerns critical growth pivot points, novel technical convergences, or transformative business models within voluminous data and volatile environments.

Trend Prediction	Systematically tracks frontier technologies and theoretical shifts to develop dynamic forecasting capabilities, enabling precise anticipation of technological trajectories and future mainstream paradigms.
Resource Integration	Identifies, acquires, and optimally allocates strategic assets to catalyze the achievement of research and commercial milestones.

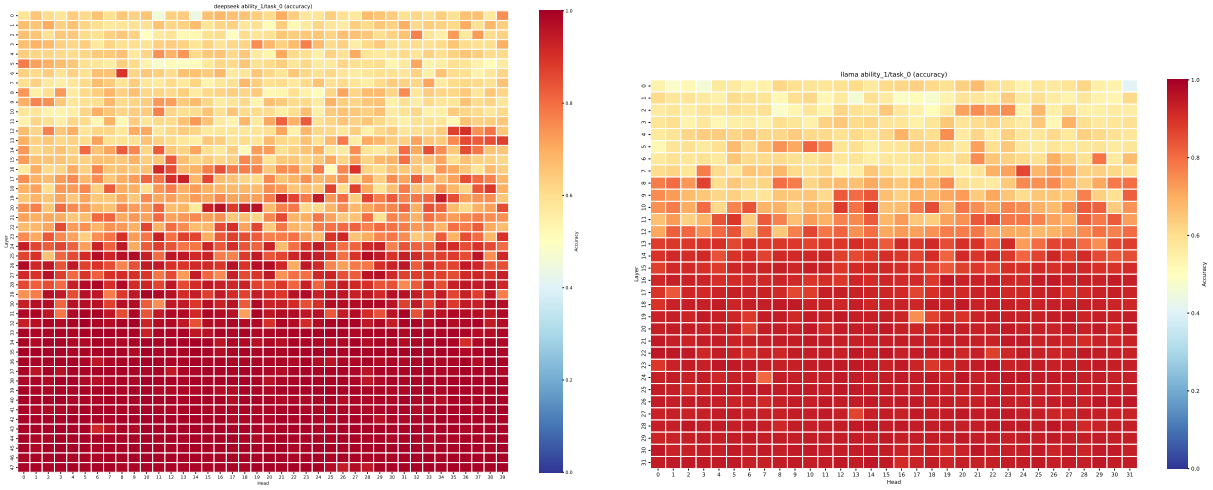
Table A1: Detailed description of 4 primary dimensions and 15 secondary dimensions in competence

Model	Subset	N	Score 0-1	Score 2	Score 3	Score 4
Qwen	Original	3990	258	2252	1474	6
	Augmented	8745	2192	2565	1558	2430
LLaMA	Original	3990	178	1140	2338	334
	Augmented	8745	1580	1919	2736	2510
DeepSeek	Original	3990	103	1898	1981	8
	Augmented	8745	1213	2995	3135	1402

Table A2: Score distribution across different LLMs for Original and Augmented subsets. The empirical data (Original) exhibits severe label imbalance, with samples heavily skewed toward specific score tiers, whereas the augmented data provides a more balanced coverage across all competency levels.

Task Name	deepseek	llama	qwen
Ethical Commitment	0.637*	0.633*	0.768**
Decisiveness	0.368	0.385	0.563*
Innovative Thinking	0.757**	0.637*	0.798***
National Commitment	0.456	0.630*	0.535*
Technical Problem-solving	0.700**	0.682**	0.717**
Opportunity Insight	0.500*	0.721**	0.475*
Communication & Influence	0.571*	0.602*	0.675**
Social Responsibility	0.696**	0.633*	0.743**
Organizational	0.607*	0.587*	0.756**
Resource Integration	0.581*	0.752**	0.747**
Trend Prediction	0.695**	0.786**	0.761**
Iterative Improvement	0.693**	0.748**	0.721**
Problem Decomposition	0.687**	0.649**	0.724**
Leadership Vision	0.701**	0.568*	0.636**
Risk Identification	0.558*	0.588*	0.601*

Table A3: Kendall's W between models and experts across 15 tasks ($p < 0.1 : *, p < 0.05 : **, p < 0.01 : ***$)



DeepSeek Llama

Figure A1: Linear probing results for Deepseek and Llama on $Ability_1/Task_0$ (Innovative Thinking).

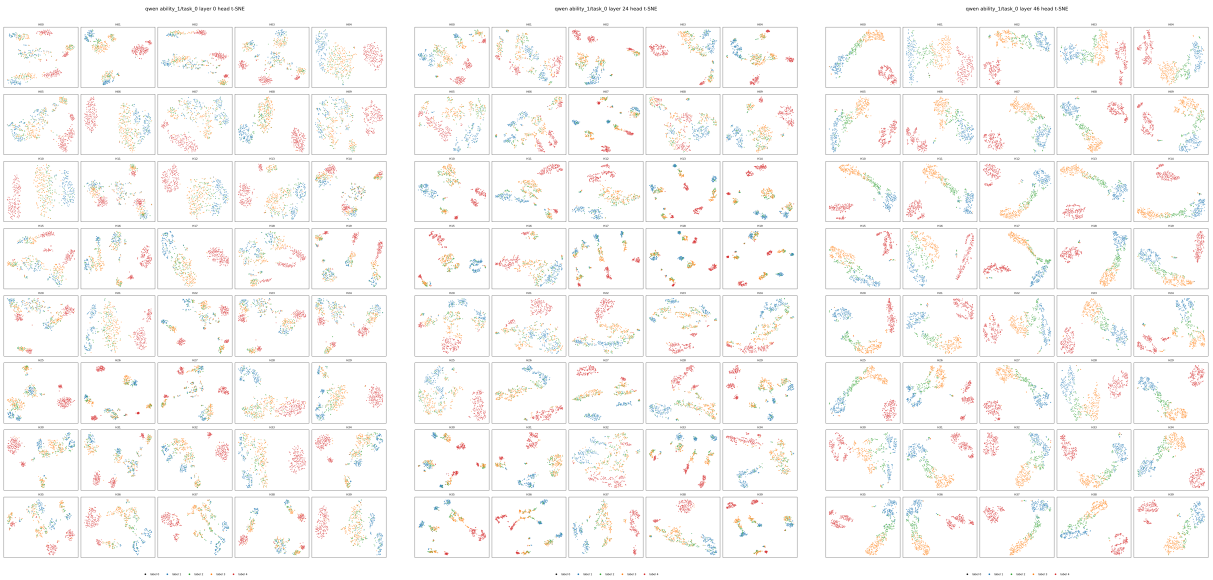
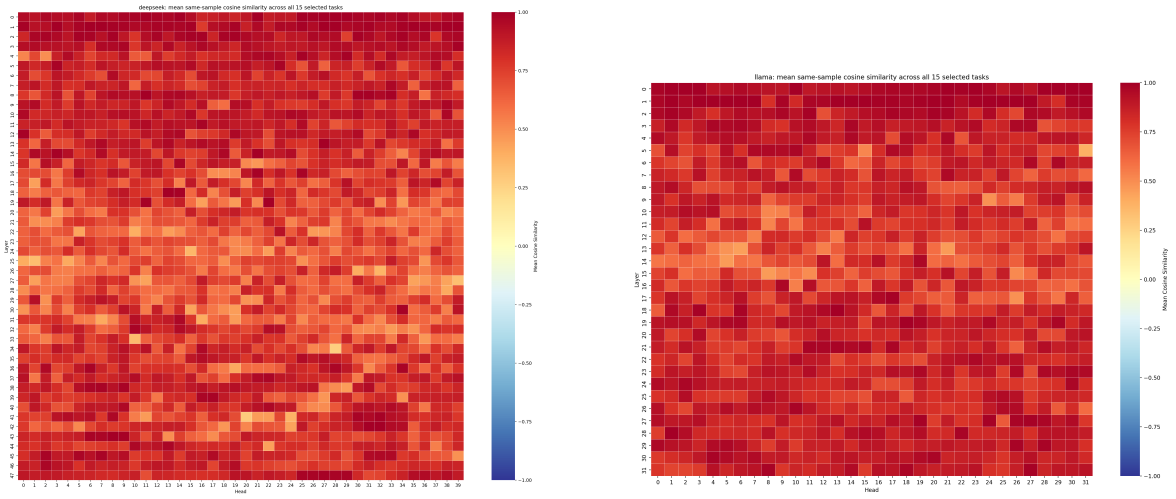
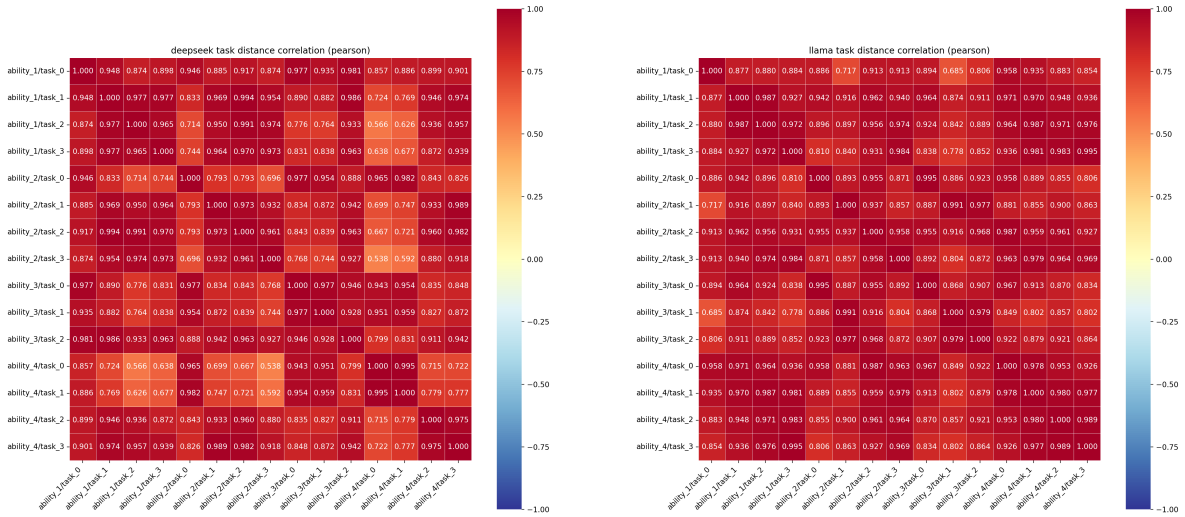


Figure A2: t-SNE clustering results of all 4-class samples on layer 0, layer 24, and layer 46 for $Ability_1/Task_0$ (Innovative Thinking).



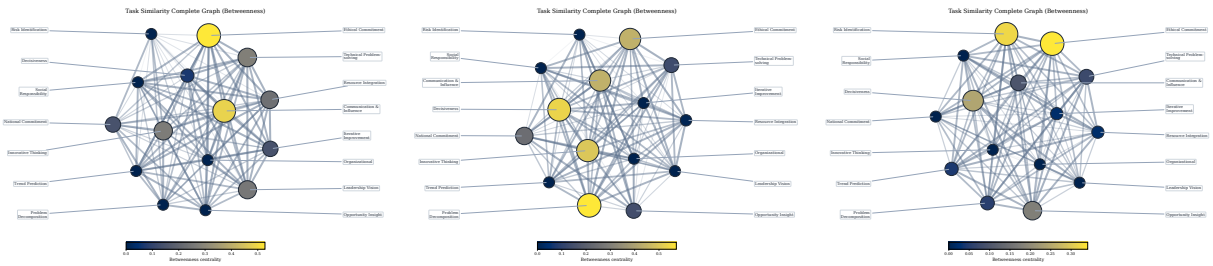
DeepSeek Llama

Figure A3: Heatmap of average cosine similarity among all 15 tasks on identical sample. X-axis denotes different heads from same layer and Y-axis denotes different layers.



DeepSeek Llama

Figure A4: Representation Similarity results for Deepseek and Llama.



Qwen DeepSeek Llama

Figure A5: A network graph visualizes the relationships between all 15 tasks in 3 models. Each node represents a specific capability while connecting edges denoting their structural similarities. The fill color of each node indicates its betweenness centrality.

Model	Task	Self (H=32, $\|A\|$=100)	Self (H=32, $\|A\|$=50)	Self (H=16, $\|A\|$=100)	Central (H=16, $\|A\|$=100)	Central (H=32, $\|A\|$=100)
Qwen	IT	+0.082 / -0.143	+0.116 / -0.048	+0.075 / -0.071	+0.116 / -0.036	+0.027 / -0.024
	PD	+0.194 / -0.065	+0.226 / -0.041	+0.194 / -0.095	+0.129 / -0.059	+0.129 / -0.083
	TPS	+0.152 / -0.177	+0.126 / -0.098	+0.212 / -0.213	+0.113 / -0.134	+0.026 / -0.079
	II	-0.084 / -0.030	-0.006 / -0.006	-0.013 / -0.006	+0.019 / -0.012	-0.136 / +0.000
	NC	+0.179 / -0.018	+0.154 / -0.024	+0.115 / -0.024	+0.154 / -0.024	-0.026 / -0.024
	LV	+0.048 / -0.060	+0.027 / -0.036	+0.068 / -0.066	-0.014 / -0.042	-0.136 / -0.036
	EC	-0.077 / -0.145	+0.021 / -0.050	+0.077 / -0.088	NA / NA	NA / NA
	SR	+0.053 / -0.012	+0.053 / -0.006	+0.060 / +0.000	+0.040 / +0.000	-0.066 / +0.000
	CI	+0.124 / -0.317	+0.178 / -0.224	+0.194 / -0.273	+0.101 / -0.099	+0.023 / -0.081
	Org	+0.082 / -0.189	+0.053 / -0.104	+0.088 / -0.128	-0.088 / -0.067	-0.265 / -0.030
	D	+0.048 / -0.006	+0.042 / -0.006	-0.018 / +0.000	-0.006 / -0.012	-0.036 / +0.000
	RId	+0.233 / -0.079	+0.302 / -0.142	+0.233 / -0.118	+0.189 / -0.157	+0.063 / -0.197
	OI	+0.176 / -0.172	+0.169 / -0.115	+0.155 / -0.083	+0.128 / -0.045	+0.061 / -0.025
	TP	+0.202 / -0.018	+0.132 / -0.012	+0.136 / -0.055	+0.054 / -0.037	-0.109 / -0.037
	RInt	+0.006 / -0.173	+0.062 / -0.025	+0.056 / -0.012	+0.099 / -0.006	+0.019 / -0.012

Table A4: Ablation for hyperparameters used in steering, where H represents for number of heads in use and $|A|$ represent the α factor in steering, where left side is score change when operating $+\alpha$ on low-score samples, while right side is operating $-\alpha$ on high-score samples.