Layer-Wise Category Structure in Large Language Models: A Cross-Architecture Analysis of Feature Decodability

Guus Bouwens [1]

Abstract

We study where category-relevant information becomes linearly decodable across layers in four instruction-tuned language models. Using residual-stream activations from 215 prompts spanning 16 task categories, we train 128 layer-wise linear probes and compare coarse early, middle, and late processing trends across architectures. We find a consistent broad representational organization in which some categories (for example, spatial navigation and logical reasoning) become decodable earlier than others, alongside meaningful architecture-specific differences in late-layer behavior such as Mistral-7B's late accuracy drop and Llama-8B's stronger confidence concentration. To keep the paper at a descriptive level, we emphasize phase-bins rather than exact layer rankings: with 16 categories, the current design has 80% power only for cross-model rank correlations of about $\rho \approx 0.65$, and fine-grained orderings are sensitive to preprocessing choices. In particular, rank agreement with the default z-score pipeline drops to $\rho = 0.287$ with raw activations and $\rho = 0.101$ with per-sample L2 normalization. Sparse autoencoder analyses provide a complementary unsupervised-style cross-check of category-selective structure, but we do not claim a localized causal mechanism. The resulting contribution is a calibrated map of broad layer-wise category structure, together with explicit boundaries on what can and cannot be concluded from the current four-model, 215-prompt dataset.

1. Introduction

Large language models can solve many tasks, but we still know little about how they organize information inside their layers. In this paper, we ask a simple question: *at which layer does each cognitive category become easy to decode from internal activations?*

Earlier probe studies (Belinkov et al., 2017; Conneau et al., 2018; Hewitt & Manning, 2019) mainly focused on language features such as syntax and semantics. Our goal is broader. We test whether layer-wise patterns for cognitive categories repeat across different model families.

To answer this, we run a layer-wise probing analysis on four models with different sizes (1.5B to 8B parameters) and depths (28 to 36 layers). We focus on three questions:

1. **Feature decodability dynamics**: At which layer does each task category become reliably decodable, and how does this vary by model family?
2. **Information flow**: Where do the largest increases in category separability occur, and do late layers refine or degrade already decodable structure?
3. **Cross-architecture comparison**: Which coarse early, middle, and late trends repeat across models, and which details remain architecture-specific?

Our contributions are:

- A reproducible framework for layer-wise category analysis using 128 probes, four models, and 16 task categories.
- Evidence for a shared coarse early-to-late representational organization across model families, with early categories such as Spatial Navigation and Logical Reasoning more consistent than later categories such as Spatial Reasoning and Executive Function.
- A calibrated account of the limits of fine-grained ranking claims: exact category order depends on prompt formulation, preprocessing, decoder family, and statistical power, so the strongest supported conclusions are phase-level rather than layer-exact.
- A multi-view descriptive analysis that combines linear probes, entropy and gain dynamics, supervised error-profile clustering, lexical-confound checks, and sparse

[1]Independent Researcher, San Diego, California, USA. Correspondence to: Guus Bouwens <guus@bouwens.nl>.

Mechanistic Interpretability Workshop at the 43rd International Conference on Machine Learning, Seoul, South Korea, 2026. Copyright 2026 by the author(s).

autoencoder feature analyses without claiming a localized causal mechanism.

2. Methods

2.1. Models and Activation Extraction

We analyze four instruction-tuned LLMs spanning a range of sizes and architectural families (Table 1). For each model, we extract residual stream activations at every transformer layer for 215 questions spanning 16 cognitive categories. Activations are extracted at the last token position after processing the full input.

Table 1. Models analyzed. All activations extracted from residual stream at each layer.

Model	Params	Layers	Hidden Dim	Architecture
DeepSeek-1.5B	1.5B	28	1536	Qwen-distilled
Qwen-4B	4.0B	36	2560	Qwen
Llama-8B	8.0B	32	4096	Llama
Mistral-7B	7.0B	32	4096	Mistral

2.2. Cognitive Categories

We probe 16 cognitive categories: Control, Creative Writing, Decision Making, Emotion, Executive Function, Language, Logical Reasoning, Math, Motor Planning, Pattern Recognition, Social Intelligence, Spatial Navigation, Spatial Reasoning, Temporal Processing, Vision, and Working Memory. The dataset has 215 questions in total, with 10–16 questions per category.

2.3. Probing Classifier Protocol

For each model layer, we train a logistic regression probe that predicts category labels from normalized activations. Across all models, this gives 128 model-layer probe settings. We use 5-fold stratified cross-validation with a fixed seed.

Normalization. Each layer is normalized to mean 0 and standard deviation 1 before training.

Evaluation. We report 16-way accuracy, per-category one-vs-rest accuracy, prediction entropy, and confusion matrices. Random guessing in a 16-class task is $1/16 = 6.25\%$.

2.4. Feature Decodability Metrics

We summarize each category by a decodability-onset layer, onset speed, and phase contributions. The onset layer is the earliest layer l^* at which per-category accuracy exceeds 70% of the range between baseline (layer 0) and peak accuracy:

$$l^* = \min\{l : a_l \geq a_0 + 0.7 \cdot (a_{\max} - a_0)\} \tag{1}$$

We use this threshold as a descriptive convention rather than a claim about a uniquely privileged layer boundary. Existing sweeps at 60%, 70%, and 80%, together with smoothed windows, preserve the same broad early-versus-late ordering, so the paper emphasizes phase bins over exact single-layer rankings.

Onset Speed. The maximum single-layer accuracy increase:

$$s = \max_l (a_{l+1} - a_l) \tag{2}$$

Phase Contributions. We decompose accuracy gain into three equal-depth phases: early (0–$\frac{1}{3}$), mid ($\frac{1}{3}$–$\frac{2}{3}$), and late ($\frac{2}{3}$–1) layers.

2.5. Information Dynamics

Information Gain. Layer-wise accuracy differences $\Delta a_l = a_{l+1} - a_l$ per category.

Prediction Entropy. Shannon entropy of the classifier's probability distribution over categories:

$$H_l = -\frac{1}{N} \sum_{i=1}^{N} \sum_{c=1}^{16} p_{i,c}^{(l)} \log_2 p_{i,c}^{(l)} \tag{3}$$

where $p_{i,c}^{(l)}$ is the predicted probability of sample i belonging to class c at layer l. Maximum entropy is $\log_2(16) = 4.0$ bits.

Cross-Layer Consistency. The fraction of samples receiving the same predicted label at consecutive layers.

2.6. Robustness Checks

We compute 95% bootstrap confidence intervals (1000 resamples) for peak accuracy and onset layers, and we test phase differences with paired Wilcoxon tests plus Benjamini–Hochberg correction. To bound the scope of the descriptive claims, we also examine macro-F1, macro balanced accuracy, one-vs-rest AUROC, preprocessing ablations (raw activations, layernorm-style normalization, and per-sample L2 normalization), alternative linear decoders (ridge and linear-hinge), and negative controls including label permutation, random projection, and shuffled-layer baselines. These checks are used to characterize stability and failure modes rather than to establish a causal mechanism.

2.7. Sparse Autoencoder Feature Discovery

To complement linear probes, we train sparse autoencoders (SAEs) on residual activations for each model. The reported SAE analyses use expansion factors of 4x and 8x, L1 sparsity coefficient $\lambda = 0.01$, learning rate 10^{-4}, zero weight decay, dead-neuron tracking with resampling during training, decoder-column normalization, and early stopping over at most 100 epochs. The encoder nonlinearity (ReLU) al-

lows SAE features to capture nonlinear structure that linear probes may miss.

We then analyze the 4x SAEs as the primary feature-discovery setting: (1) feature specialization by category and layer, (2) dominant-category counts per model, and (3) cross-model feature alignment using cosine similarity of category-selectivity profiles with reciprocal nearest-neighbor matching.

3. Results

3.1. Overall Classification Performance

All four models achieve peak 16-way classification accuracy far above the 6.25% random baseline, confirming that cognitive category information is encoded in residual stream activations (Table 2).

Table 2. Peak classification performance with 95% bootstrap confidence intervals.

Model	Peak	95% CI	Best Layer	Rel. Depth	× Rand.
DeepSeek-1.5B	43.3%	[39.1%, 50.7%]	21/28	0.75	6.9×
Qwen-4B	38.1%	[34.4%, 46.0%]	35/36	0.97	6.1×
Llama-8B	52.1%	[47.9%, 59.5%]	18/32	0.56	8.3×
Mistral-7B	58.1%	[52.1%, 65.1%]	16/32	0.50	9.3×

Mistral-7B achieves the highest peak accuracy (58.1%) despite having fewer parameters than Llama-8B (52.1%). The 7–8B models significantly outperform smaller models, with non-overlapping bootstrap confidence intervals. Notably, the best layer varies dramatically: Qwen-4B peaks at its penultimate layer (35/36), while Mistral-7B peaks at its midpoint (16/32).

Figure 1 makes this pattern concrete. In the top row heatmaps, bright regions appear earlier for some categories and later for others, showing that category information does not appear uniformly. In the bottom-left accuracy panel, all models rise quickly early, then diverge in late layers. In the bottom-right emergence timing panel, warmer colors for late categories (for example Pattern Recognition and Executive Function) visually confirm the late-emergence tendency reported in Table 3.

3.2. Broad Early-to-Late Representational Organization

Table 3 summarizes the average onset ranking across the four architectures. The main descriptive pattern is coarse rather than exact: categories such as Spatial Navigation and Logical Reasoning tend to become decodable earlier, while categories such as Spatial Reasoning, Executive Function, and Pattern Recognition tend to become decodable later.

The earlier categories also show lower cross-model variance than the later categories, which is why we interpret the strongest shared signal as a broad early-versus-late organization. Figure 1 shows the same pattern visually: early categories appear in earlier bands across models, while later categories shift more with architecture and preprocessing choices.

Table 3. Category onset ranking, averaged across 4 models. Lower = earlier onset. Consensus indicates cross-model agreement (std < 0.15: High).

Rank	Category	Mean Onset	Std	Consensus
1	Spatial Navigation	0.212	0.066	High
2	Logical Reasoning	0.215	0.043	High
3	Control	0.247	0.101	High
4	Motor Planning	0.252	0.103	High
5	Vision	0.262	0.074	High
12	Creative Writing	0.367	0.221	Medium
13	Temporal Processing	0.420	0.240	Medium
14	Spatial Reasoning	0.483	0.319	Low
15	Executive Function	0.515	0.244	Medium
16	Pattern Recognition	0.532	0.180	Medium

3.3. Early-Phase Dominant Learning

All four models show a consistent pattern: the majority of category separability is established in the first third of layers (Table 4). The same trend appears in Figure 1 (bottom-left), where the steepest slope is concentrated in early depth for each model.

Table 4. Phase contributions to accuracy gain. All models are early-dominant.

Model	Early (0–$\frac{1}{3}$)	Mid ($\frac{1}{3}$–$\frac{2}{3}$)	Late ($\frac{2}{3}$–1)	Dominant
DeepSeek-1.5B	+0.0215	+0.0006	+0.0081	Early
Qwen-4B	+0.0279	−0.0017	+0.0052	Early
Llama-8B	+0.0337	+0.0035	+0.0000	Early
Mistral-7B	+0.0302	+0.0099	−0.0140	Early

Late-layer behavior differs across architectures:

- **Mistral-7B** shows a negative late-phase contribution (-1.4%), so separability drops in final layers.
- **Llama-8B** shows exactly zero late contribution, with all discriminative work completed by mid-network.
- **Qwen-4B** shows mild mid-phase regression (-0.17%) followed by late-phase recovery.
- **DeepSeek-1.5B** is the only model showing positive contribution across all three phases.

3.4. Robustness Boundaries and Surface Confounds

Alternative metrics preserve the high-level model ordering: peak macro-F1 is highest for Mistral-7B (0.575), followed by Llama-8B (0.514), DeepSeek-1.5B (0.431), and Qwen-4B (0.369); balanced accuracy and AUROC show the same ranking. Early-phase gains exceed both mid and late gains

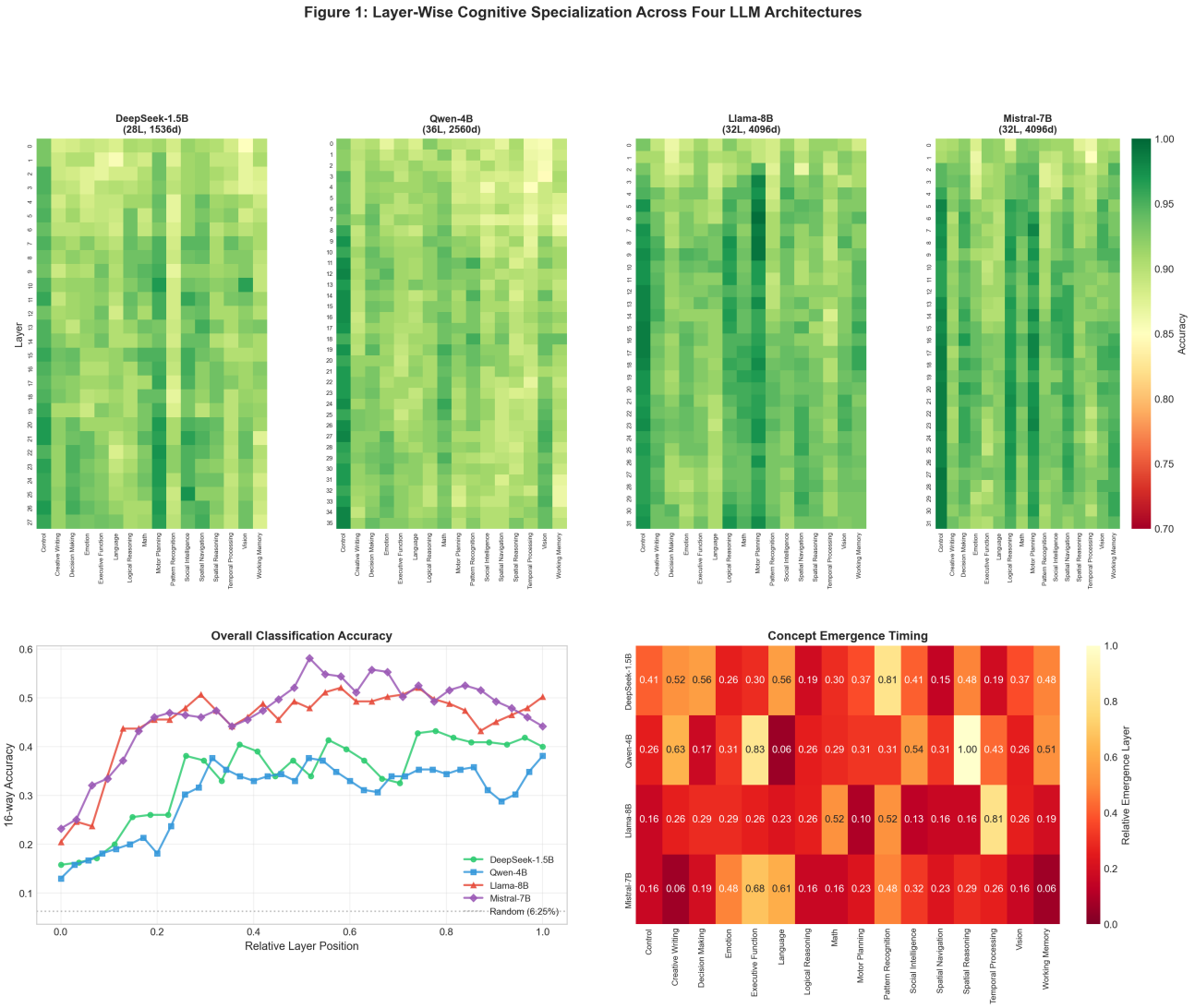


Figure 1. Overview: competency heatmaps (layer × category) for all four models, overall accuracy curves, and emergence timing heatmap.

($p_{\mathrm{BH}} = 1.15 \times 10^{-6}$ and 2.87×10^{-8}), but the exact 16-category ordering is statistically fragile. A global permutation test on pairwise rank agreement is not significant ($p = 0.526$), and with 16 categories the current design has 80% power only for cross-model rank correlations of about $\rho \approx 0.65$. Observed pairwise correlations (-0.18 to 0.15) therefore rule out strong rank agreement but not modest agreement.

The broad early/mid/late pattern is more stable than exact layer indices. Under preprocessing ablations, mean rank agreement with the default z-score pipeline drops to $\rho = 0.287$ for raw activations and $\rho = 0.101$ for per-sample L2 normalization, while a layernorm-style transform preserves moderate agreement ($\rho = 0.497$). Decoder choice also changes fine-grained rankings: ridge and linear-hinge decoders yield only moderate mean rank agreement with logistic probes ($\rho = 0.277$ and $\rho = 0.345$, respectively). We therefore interpret the main result as coarse representational organization, not exact single-layer timing. The 70% onset threshold is a descriptive convention rather than a tuned claim; threshold sweeps at 60% and 80%, together with smoothed windows, preserve broadly similar early-versus-late orderings (mean rank agreement 0.70–0.81 versus the default).

Surface-text baselines show that lexical cues matter but do not fully explain the probe results. A TF–IDF classifier reaches 0.386 accuracy, yet probe confusion structure diverges meaningfully from surface similarity: the mean top-10 confusion overlap with the surface baseline is only 1.75/10, and the mean off-diagonal Pearson correlation is

0.308. Negative controls behave as expected: label permutation collapses performance near random, and random projection or shuffled-layer baselines reduce performance relative to the true residual features. Together these checks support a narrow claim: the probes recover non-trivial internal structure, but only at the level of decodable correlates rather than causal mechanisms.

3.5. Information Dynamics

Figure 4 links four complementary views of the same process. Panel (A) tracks confidence concentration, panel (B) shows where layer-to-layer gains occur, panel (C) summarizes trajectory shape, and panel (D) reports prediction stability across adjacent layers. Reading these panels together helps separate three effects: where information appears, how strongly it appears, and whether it remains stable.

3.5.1. Critical Layers

Information gain analysis (Figure 4B) reveals that the most critical processing occurs at very different relative positions across architectures: Mistral-7B (layer 1, 3% depth), Llama-8B (layer 2, 6%), Qwen-4B (layer 7, 20%), and DeepSeek-1.5B (layer 19, 70%). Larger models concentrate their critical differentiation earlier, suggesting that greater capacity can support faster category separation.

3.5.2. Entropy Dynamics

Prediction entropy reveals qualitatively different confidence trajectories (Figure 4A):

- **Llama-8B**: The only model with substantial entropy reduction (0.41 bits), reaching minimum entropy at its final layer. This model genuinely becomes *more confident* about category identity as depth increases.
- **DeepSeek-1.5B and Qwen-4B**: Both show *increasing* entropy across layers. Accuracy still improves, which suggests improved separability can happen without confidence concentration.
- **Mistral-7B**: Nearly flat entropy trajectory (0.02 bit drop), with minimum entropy at layer 20, suggesting that Mistral maintains consistent uncertainty throughout processing.

Panel (D) adds an important complement: even when entropy remains high, cross-layer consistency can still increase. This helps explain why some models improve classification without sharp confidence concentration.

3.6. Confusion Structure

Analysis of confusion matrices at each model's best layer reveals systematic cross-category confusions (Table 5). Figure 3 shows that these errors are structured rather than random: diagonal entries remain strongest, while a small number of off-diagonal pairs recur across models. Notably, the confusion between **Social Intelligence** → **Emotion** appears in all four models, suggesting a fundamental representational overlap between social and emotional processing in transformer architectures.

Table 5. Most common cross-category confusions. "Shared" indicates presence in $\geq 3/4$ models.

True Category	Predicted As	Models	Shared?
Social Intelligence	Emotion	4/4	✓
Creative Writing	Language	3/4	✓
Language	Creative Writing	3/4	✓
Logical Reasoning	Math	2/4	
Executive Function	Various	3/4	✓

The bidirectional Creative Writing ↔ Language confusion is particularly notable. Classifiers confuse these categories in both directions, which suggests strong representational overlap. Figure 3 shows the full confusion structure per model.

Prompt-form controls qualify this picture without removing it. A TF-IDF surface baseline reaches 0.386 accuracy and 0.351 macro-F1, so wording alone carries real category signal. However, probe confusions only partially overlap with surface confusions: the mean off-diagonal Pearson correlation is 0.308, the mean Spearman correlation is 0.179, and the mean top-10 pair overlap is 1.75. Recurrent probe confusions therefore reflect both prompt-level cues and deeper representational overlap.

3.7. Auxiliary Cross-Model Structure

Pairwise onset correlations are near zero ($\rho = -0.18$ to 0.15), even though full layer-accuracy profiles correlate moderately ($r = 0.38$–0.60). This again supports the same calibrated conclusion: coarse early-versus-late structure is more stable than exact onset timing.

Taxonomy controls and SAE cross-checks are consistent with that picture. Merged-taxonomy controls preserve moderate rank agreement ($\rho = 0.535$, 0.677, and 0.764 for three label-merging schemes), and the best-layer confusion map is best interpreted as a *supervised error-profile taxonomy* rather than a fully unsupervised taxonomy. The 4x SAE inventories remain fully active across all four models with strong category-selective features, providing a complementary feature-level check on the probe-based findings. Appendix G and Appendix F provide the full clustering and SAE diagnostics.

4. Discussion

4.1. Broad Early-to-Late Organization

One central finding is that category information follows a similar coarse order across four different LLM architectures. Spatial navigation and logical reasoning tend to separate earlier; pattern recognition and executive function tend to separate later. This pattern is visible both numerically (Table 3) and visually (Figure 1, onset panel), but it should be read as a phase-level regularity rather than a stable exact ranking.

At the same time, our permutation test shows that full 16-category rank agreement is not statistically significant at this sample size. The strongest evidence is for broad early-versus-late tendencies.

The newer control suite makes this interpretation more precise. Broad order is more stable under merged taxonomies and balanced weighting than under alternative preprocessing or different decoder families. The strongest reviewer-facing claim should therefore stay at the level of broad representational organization, not exact layer timing.

4.2. Architecture-Specific Late-Layer Behavior

Late-layer dynamics differ across models. Mistral shows forgetting, Llama plateaus, Qwen shows mid-phase regression with later recovery, and DeepSeek keeps improving. These differences are visible in Figure 1 (accuracy curves) and Table 4 (signed phase contributions). Together they suggest each architecture balances representation quality and output behavior in its own way.

Mistral-7B's late-layer accuracy degradation (-1.4%) suggests that final layers may be optimized for output behavior, with less category-separable information preserved. For applications that use internal states directly (retrieval, probing, steering), mid-layers may provide stronger signals.

4.3. Limitations

Our work has several limitations. First, with 16 categories and 215 prompts, the design has 80% power only for cross-model rank correlations of about $\rho \approx 0.65$ or larger, so it is underpowered for fine-grained ranking claims. Second, the broad ordering depends on the default z-score pipeline: mean rank agreement drops to $\rho = 0.287$ with raw activations and $\rho = 0.101$ with per-sample L2 normalization, and exact rankings also shift under ridge and linear-hinge decoders. Third, prompt formulation remains a factor; the 62-prompt auxiliary paraphrase set is useful for transparency about wording sensitivity, but it is too small to support strong invariance claims. Finally, linear probes reveal only decodable correlates of the categories. Because this paper does not rely on convergent feature-level interventions, it does not establish the causal mechanisms that drive model behavior.

4.4. Scope of Claims

This workshop paper stays strictly at Tier 1 descriptive robustness. The supported claim is that broad layer-wise category structure and coarse phase-bin trends are reproducible across several analyses in four models. The paper does not claim that probes identify the mechanisms that drive behavior, and it does not claim that the exact ranking is stable across architectures or prompt formulations.

5. Related Work

Probing classifiers and layer-wise analyses have long been used to study what information is decodable from neural representations and where it concentrates across depth (Belinkov et al., 2017; Conneau et al., 2018; Hewitt & Manning, 2019; Jawahar et al., 2019; Tenney et al., 2019). Mechanistic interpretability and cross-model comparison work further motivate the use of SAE cross-checks, onset correlations, and profile-similarity analyses (Elhage et al., 2021; Geva et al., 2021; Cunningham et al., 2023; Bricken et al., 2023; Kriegeskorte et al., 2008; Bansal et al., 2021). Our contribution is to apply those descriptive tools to a manually designed set of 16 task categories while explicitly bounding the scope of the resulting claims.

6. Conclusion

We have presented a layer-wise analysis of category structure in four LLM architectures. The most defensible result is a descriptive one: broad early-to-late regularities, architecture-specific late-layer dynamics, and stable supervised error-profile overlaps can be observed in the current dataset, while exact onset rankings remain sensitive to prompt wording, preprocessing, decoder choice, and limited statistical power. SAE analyses provide a complementary feature-level cross-check, but not a localized causal account. The paper should therefore be read as a calibrated map of broad representational organization rather than proof of an architecture-invariant hierarchy or mechanism.

Future work. The most valuable next steps are now narrower: (1) enlarge the paraphrase set with more structurally diverse prompt families, (2) run aligned token-position or feature-level interventions that can test specific mechanisms, and (3) connect those interventions more directly to sparse autoencoder feature groups.

References

Bansal, Y., Nakkiran, P., and Barak, B. Revisiting model stitching to compare neural representations. In *NeurIPS*,

2021.

Belinkov, Y., Durrani, N., Dalvi, F., Sajjad, H., and Glass, J. What do neural machine translation models learn about morphology? In *Proceedings of ACL*, 2017.

Bricken, T., Templeton, A., Batson, J., Chen, B., Jermyn, A., Conerly, T., Turner, N., Anil, C., Denison, C., Askell, A., et al. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023.

Conneau, A., Kruszewski, G., Lample, G., Barrault, L., and Baroni, M. What you can cram into a single vector: Probing sentence embeddings for linguistic properties. In *Proceedings of ACL*, 2018.

Cunningham, H., Ewart, A., Riggs, L., Huben, R., and Sharkey, L. Sparse autoencoders find highly interpretable features in language models. In *Proceedings of ICLR*, 2023.

Elhage, N., Nanda, N., Olsson, C., Henighan, T., Joseph, N., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021.

Geva, M., Schuster, R., Berant, J., and Levy, O. Transformer feed-forward layers are key-value memories. In *Proceedings of EMNLP*, 2021.

Hewitt, J. and Manning, C. D. A structural probe for finding syntax in word representations. In *Proceedings of NAACL*, 2019.

Jawahar, G., Sagot, B., and Seddah, D. What does BERT learn about the structure of language? In *Proceedings of ACL*, 2019.

Kriegeskorte, N., Mur, M., and Bandettini, P. Representational similarity analysis – connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2(4), 2008.

Tenney, I., Das, D., and Pavlick, E. BERT rediscovers the classical NLP pipeline. In *Proceedings of ACL*, 2019.

A. Methodological Transparency

A.1. Category Design and the Reasoning Scaffold

The 16 reported categories were not chosen ad hoc. The scaffold file `reasoning_scaffold.json` in `results/category_clustering/` specifies nine design axes: `verbal_generation`, `symbolic_rules`, `social_inference`, `valuation_choice`, `spatial_transform`, `motor_sequencing`, `perception`, `temporal_ordering`, and `executive_control`. Each category is assigned axis weights, expected neighbors, and optional specialist/bridge flags. This scaffold was used as an explicit design prior for interpreting confusion structure and for checking whether nearby categories in the probe results align with intended cognitive overlap. The paper therefore studies a manually designed task taxonomy, not a latent ontology discovered from raw activations alone.

A.2. Construction of the 62-Prompt Paraphrase Set

The paraphrase scaffold in `data/replication_set/paraphrase_template.json` defined five rewrite slots per original prompt: *lexical*, *structural*, *concise*, *adversarial*, and *long_context*, with status values `todo`, `draft`, and `ready`. Across 215 original prompts this created 1,075 possible paraphrase slots. The specific replication analysis reported in this paper used only the ready items available at analysis time: 62 paraphrases total, consisting of 30 lexical rewrites and 32 structural rewrites across the 16 categories. The concise, adversarial, and long-context slots were scaffolded for future expansion but were not part of the reported 62-question evaluation. This is why the paraphrase evidence is treated here as auxiliary wording-sensitivity context rather than as a definitive robustness benchmark.

A.3. SAE Training Configuration

The SAE runs summarized in the main text were trained on residual activations with 4x and 8x expansion factors. The implementation used ReLU latents, L1 sparsity coefficient $\lambda = 0.01$, Adam with learning rate 10^{-4} and zero weight decay, decoder-column normalization, question-level train/validation/test splits, early stopping with patience 15 over at most 100 epochs, and dead-neuron monitoring with resampling during training. Training summaries for all four models are stored in `results/sae/training_summary.json`; in the held-out summaries, the reported runs show zero dead-neuron fraction at evaluation time.

B. Extended Visualization and Results

This appendix contains extended visualizations and supplementary results referenced in the main text.

B.1. Probe Family Peak Accuracy

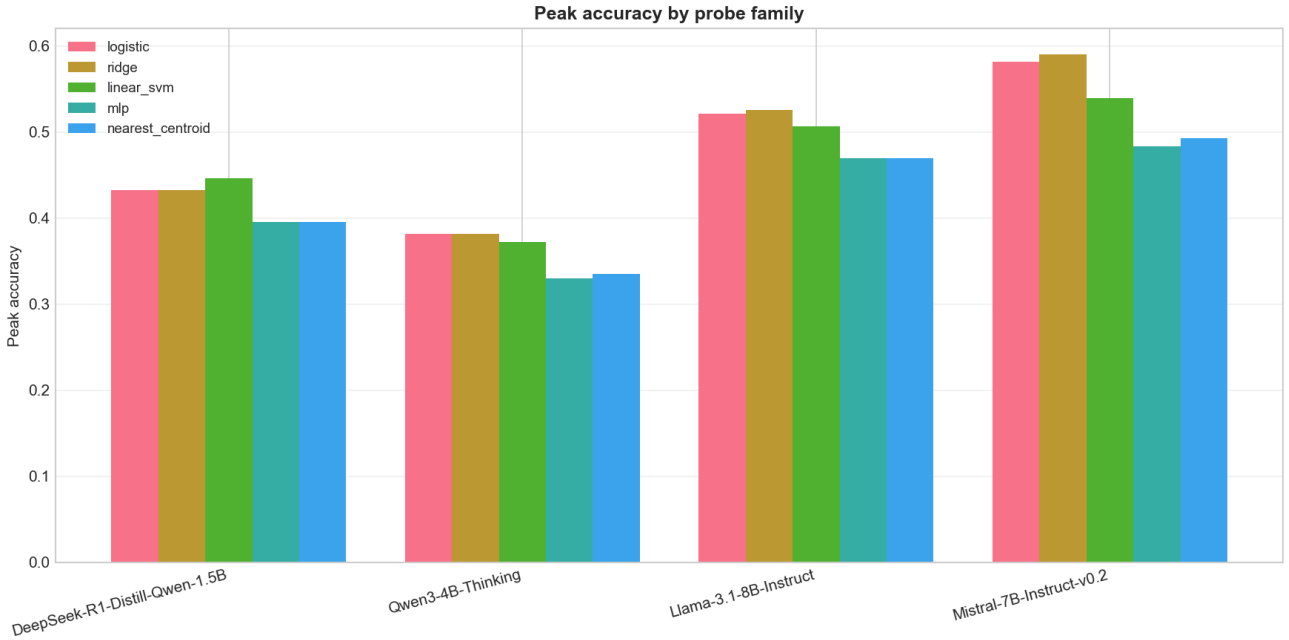


Figure 2. Peak classification accuracy across decoder families and models. Full residual-stream logistic probes consistently achieve the highest accuracy. PCA-reduced and nearest-centroid decoders retain substantial signal, confirming that category structure is not exclusively linear. SAE latent features remain above chance, providing an unsupervised cross-check. Random projections fall near chance, confirming that structure depends on the specific feature space rather than dimensionality alone.

B.2. Confusion Matrices

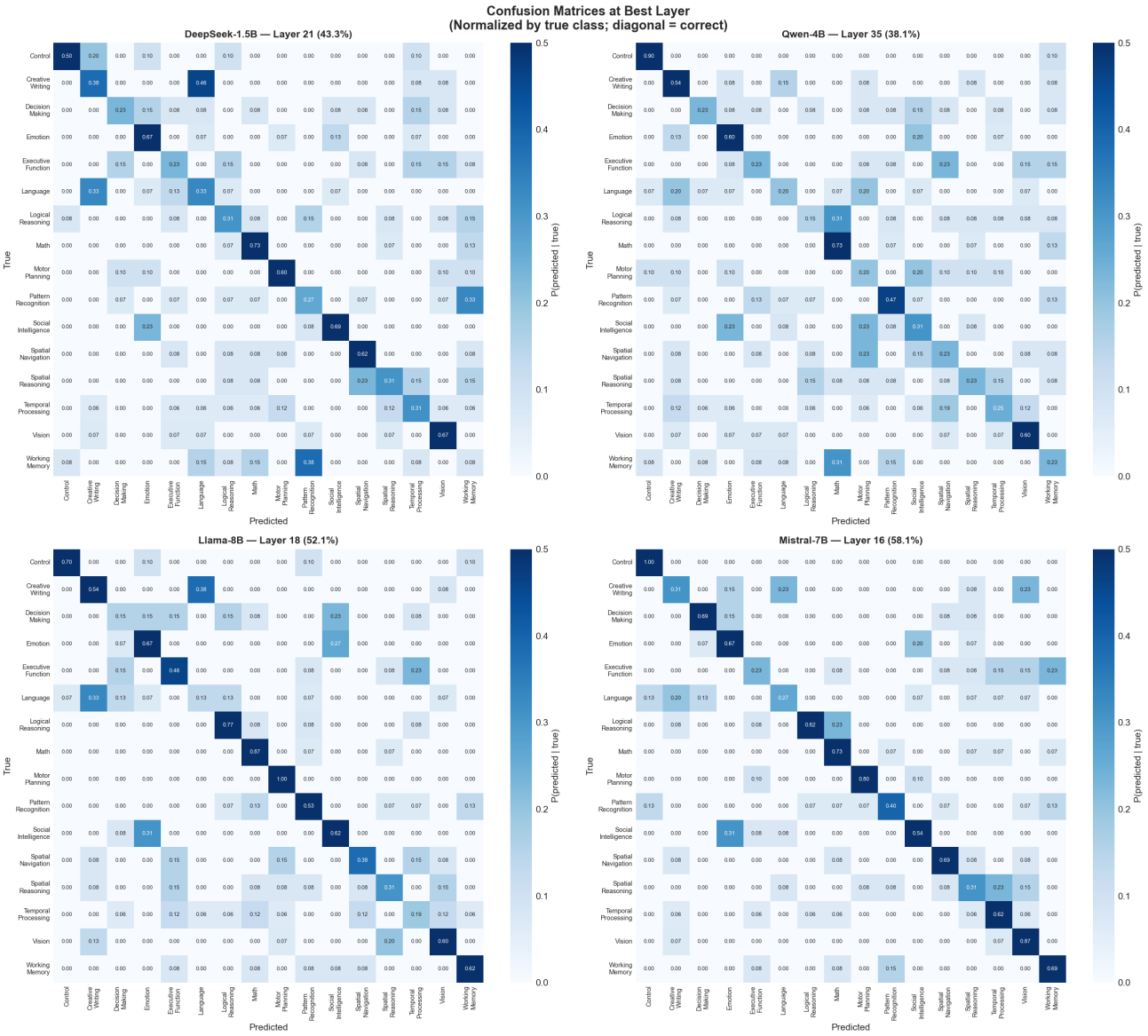


Figure 3. Confusion matrices for probe predictions (selected layers or models); rows true class, columns predicted.

B.3. Information Dynamics

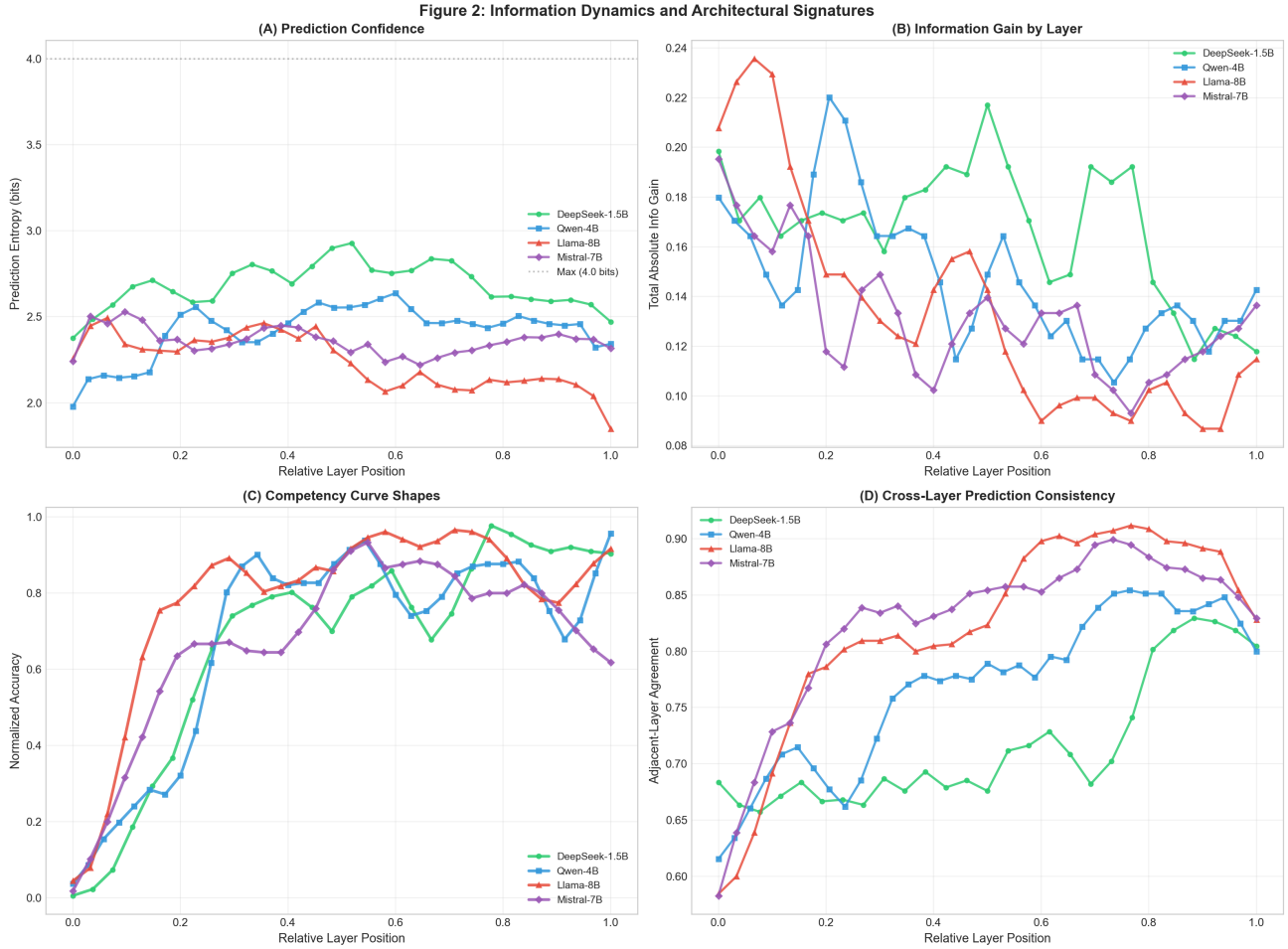


Figure 4. Information dynamics: (A) prediction entropy by layer, (B) information gain, (C) competency curve shapes, (D) cross-layer consistency.

B.4. Cross-Architecture Efficiency

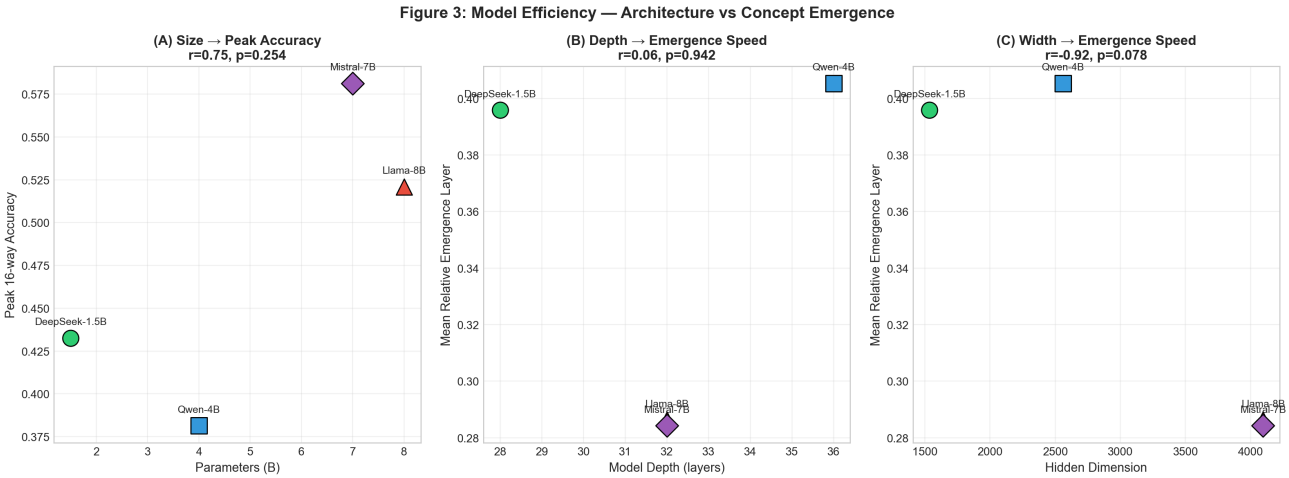


Figure 5. Efficiency vs. comprehension: onset layer and accuracy by model; architectural differences in category-encoding speed.

B.5. Representational Similarity

Figure 6. Cross-model representational similarity. **(A)** Spearman ρ for pairwise emergence timing: all model pairs fall near zero (-0.18 to 0.15), confirming that emergence *timing* is architecture-specific. **(B)** Pearson r for full layer-accuracy profiles: moderate positive correlations ($r = 0.38$–0.60), strongest for the Llama–Mistral pair ($r = 0.60$), consistent with their shared hidden dimension and design.

B.6. Category Clustering

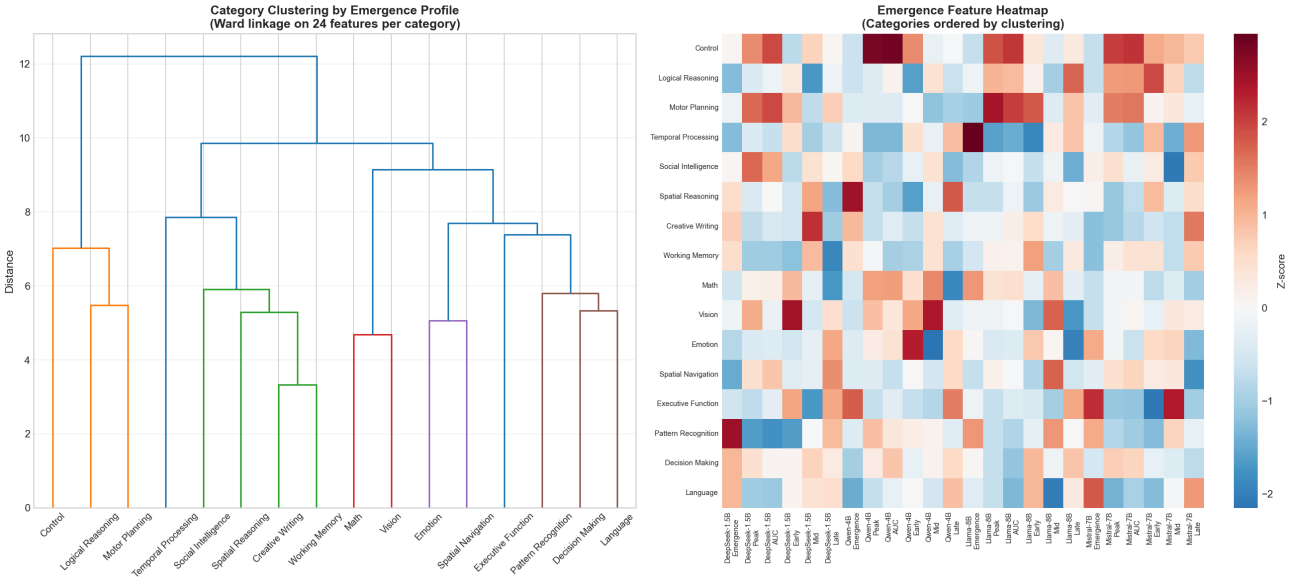
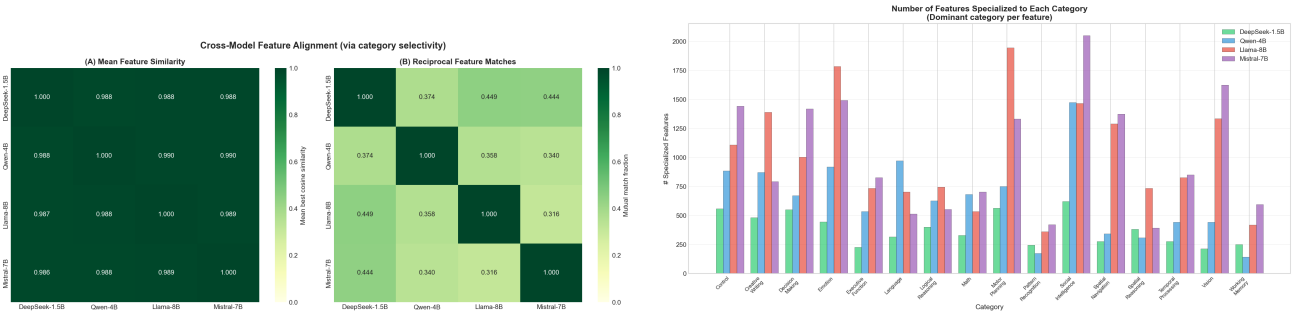


Figure 7. Hierarchical clustering (Ward linkage) of all 16 categories using 24-dimensional emergence feature vectors aggregated across models. Four natural groups emerge: early-structured (Control, Logical Reasoning, Motor Planning), late-variable (Creative Writing, Social Intelligence, Spatial Reasoning, Temporal Processing, Working Memory), unique-profile (Math, Vision), and mixed (remaining categories). Colors encode cluster assignment.

B.7. SAE-Based Unsupervised Feature Discovery



(a) Cross-model SAE feature alignment

(b) Category-specialized SAE feature counts

Figure 8. SAE-based unsupervised feature discovery. Panel (A) shows strong cross-model alignment in category-selectivity space. Panel (B) shows that all categories are represented by specialized features across architectures.

C. Full Competency Heatmaps

Full competency heatmaps for all four models are provided in Figure 1 (main text). Each cell (l, c) represents the probing accuracy for layer l on category c, with green indicating high accuracy and red indicating low accuracy.

D. Per-Model Emergence Details

Table 6. Layer specialization summary per model.

Metric	DeepSeek	Qwen	Llama	Mistral
Mean specialization score	0.038	0.049	0.042	0.042
Most specialized layer	2	34	3	31
Critical info-gain layer	19 (70%)	7 (20%)	2 (6%)	1 (3%)
Initial entropy (bits)	2.376	1.979	2.257	2.241
Final entropy (bits)	2.472	2.343	1.850	2.317
Entropy change	+0.096	+0.364	−0.407	+0.076

E. Bootstrap Confidence Intervals

Full bootstrap results (1000 resamples) for onset layers per category are shown in Figure 9. The coarse early-phase dominance and late-layer divergence patterns remain visible under resampling.

F. SAE Supplementary Diagnostics

This appendix provides additional diagnostics for feature inventory quality and training behavior. The same conclusions hold: feature dictionaries remain active, category-selective, and comparable across models.

The inventory view in Figure 10 is consistent with the main-text alignment analysis (Figure 8), providing an architectural comparison at the feature quality level in addition to category-selectivity structure.

G. Category Taxonomy from Supervised Error Profiles

Figure 11 shows a category map built from best-layer confusion profiles. The clustering step itself is unsupervised, but the profiles come from supervised probes, so this is a supervised error-profile analysis rather than a fully unsupervised embedding taxonomy. The map still recovers meaningful pairings: Creative Writing clusters with Language, Logical Reasoning with Math, and Social Intelligence with Emotion. This agreement with the supervised probe confusions (Table 5) supports the claim that confusion structure reflects representational overlap rather than purely arbitrary label assignment.

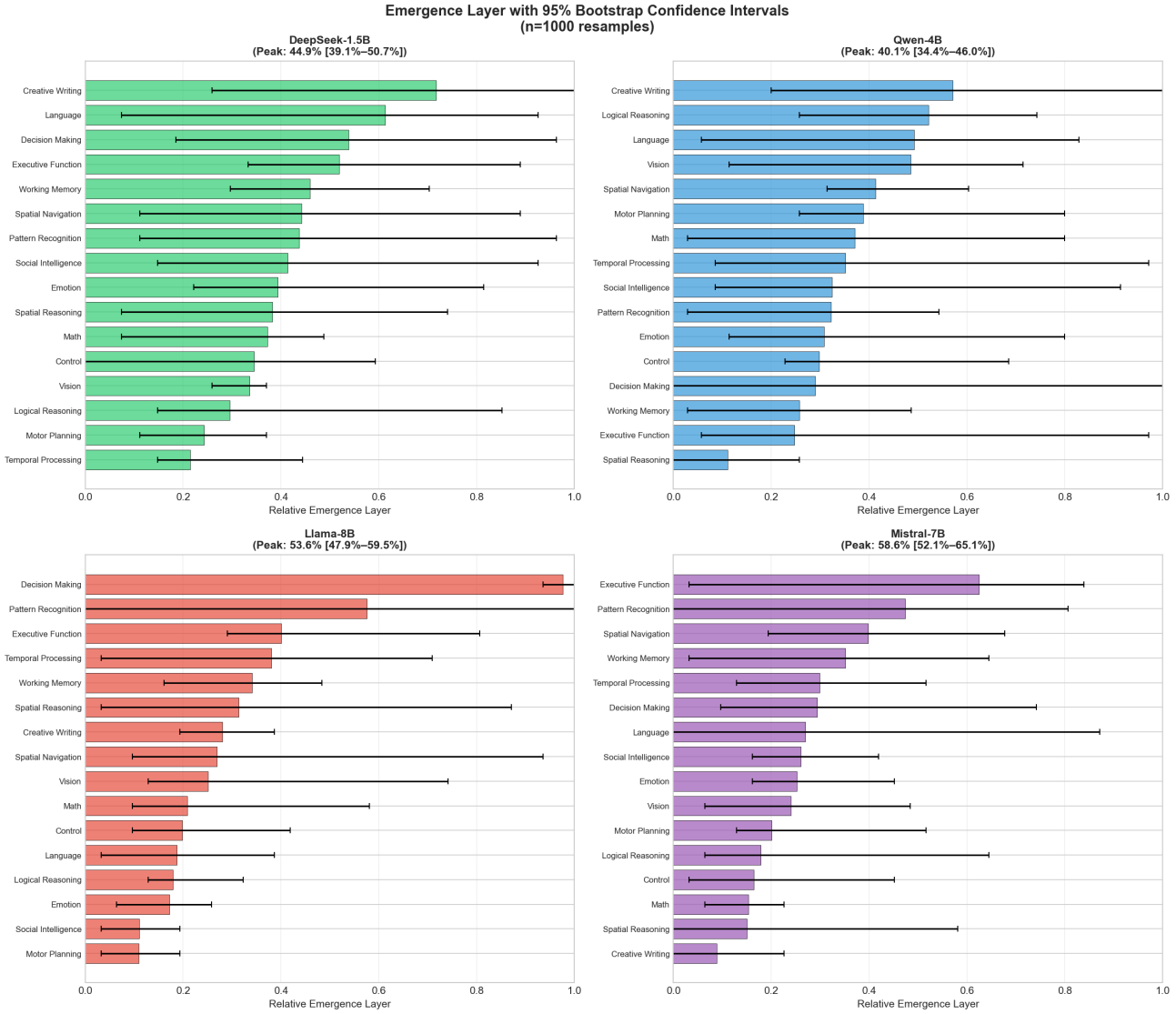


Figure 9. Bootstrap 95% confidence intervals for emergence layers; stability of emergence estimates.

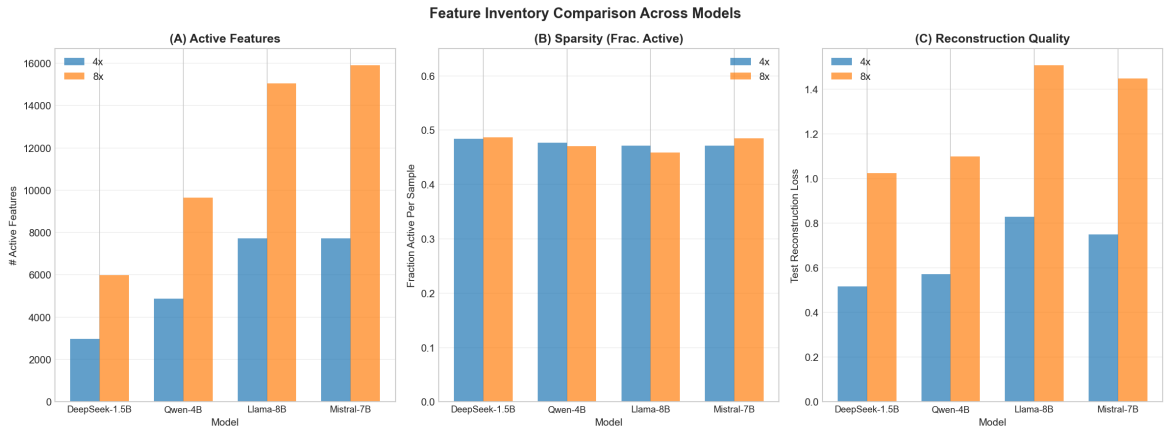


Figure 10. SAE inventory diagnostics across models and expansion factors: mean active features per sample, sparsity (fraction active), and reconstruction loss.

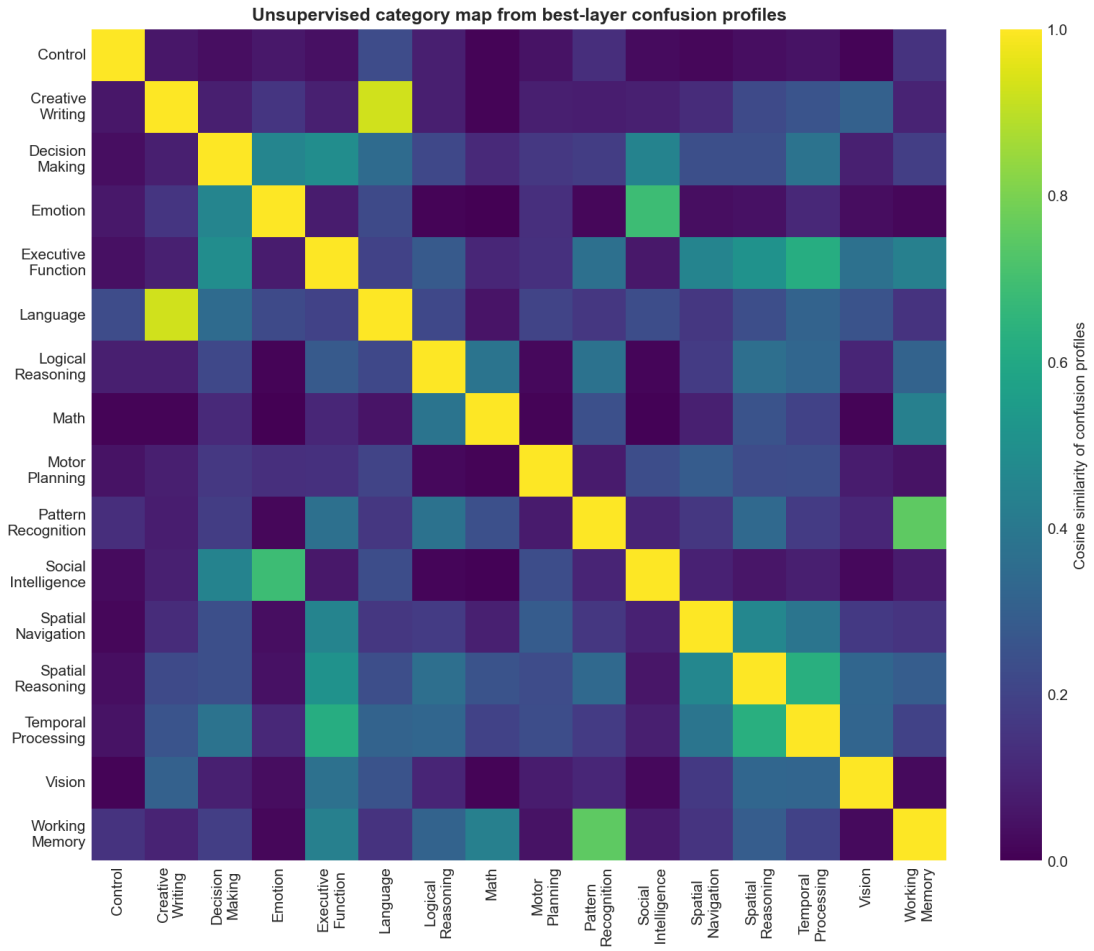


Figure 11. Category map constructed from best-layer confusion profiles across all four models. Proximity indicates similarity of supervised error profiles. Recovered pairs (Creative Writing–Language, Logical Reasoning–Math, Social Intelligence–Emotion) match the supervised confusion structure from Table 5.