Cogni Map: Real-Time Detection of Cognitive Actions in Language Models Through Linear Probing

Ivan Chulo

Harvard University
Cambridge, MA 02138 USA
ichulo@g.harvard.edu

Abstract

We present Cogni Map, a system for detecting 45 distinct cognitive actions in real-time as large language models generate text. Unlike prior work on profiling user demographics (**?**), we focus on identifying the cognitive processes exhibited in model-generated content itself—such as *analyzing*, *reconsidering*, *divergent_thinking*, and *self_questioning*. Using linear probes trained on internal activations from Gemma-3-4B, we achieve an average AUC-ROC of 0.78 across all cognitive actions. Our approach employs an optimized single-pass activation capture method that is 25× faster than sequential layer extraction, and trains binary one-vs-rest probes for each cognitive action. We demonstrate layer specialization, where different cognitive processes are best detected at different model depths, and present applications to therapy transcript analysis. This work bridges mechanistic interpretability and cognitive science, offering a new lens for understanding AI reasoning.

Introduction

As large language models (LLMs) become increasingly capable, understanding their internal reasoning processes becomes critical for safety, interpretability, and alignment (**?**). While recent work has explored profiling user attributes from conversations (**?**), little attention has been paid to identifying the *cognitive actions* exhibited by the model itself during text generation.

We introduce Cogni Map, a mechanistic interpretability tool that detects 45 cognitive actions spanning metacognitive, analytical, creative, and emotional categories. Inspired by cognitive psychology taxonomies, these actions provide a fine-grained vocabulary for describing AI "thought processes"—from *pattern_recognition* and *hypothesis_generation* to *emotional_reappraisal* and *counterfactual_reasoning*.

Our approach builds on linear probing techniques (**??**) to extract cognitive action representations from transformer activations. Unlike dashboard systems for user model transparency (**?**), we focus on model-internal cognitive processes, enabling researchers to observe "how the model thinks" rather than "who it thinks it's talking to."

Contributions: (1) A dataset of 6,750 synthetic examples across 45 cognitive actions with augmented prompting for activation capture; (2) An optimized single-pass activation capture method reducing compute by 25×; (3) Binary probes achieving 0.78 average AUC-ROC with identified layer specialization; (4) Real-world application to therapy transcript analysis demonstrating practical utility.

Methodology

Cognitive Action Taxonomy

We define 45 cognitive actions organized into four categories: **Metacognitive** (13 actions: reconsidering, updating_beliefs, meta_awareness, etc.), **Analytical** (12 actions: analyzing, evaluating, abstracting, etc.), **Creative** (6 actions: divergent_thinking, reframing, analogical_thinking, etc.), and **Emotional** (14 actions: emotional_reappraisal, emotion_perception, etc.). This taxonomy draws from cognitive psychology research on reasoning, creativity, and emotional intelligence.

Activation Capture Pipeline

Following Chen et al.'s (**?**) approach to probing internal representations, we use nnsight (**?**) to extract activations from Gemma-3-4B at layers [7, 14, 21, 27]. Crucially, we employ **augmented prompting** to create a consistent extraction point:

```
augmented_text = f"{text}\n\nThe
cognitive action being demonstrated
here is"
```

This primes the model to encode cognitive action information in the final token representation. Our key innovation is **single-pass multi-layer capture**, extracting all four layers simultaneously during one forward pass (25× speedup vs. sequential):

```
with model.trace(augmented_text):
    for layer_idx in [7, 14, 21, 27]:
        saved_states[layer_idx] =
            layers[layer_idx].output[0].save()
```

Binary Probe Training

We use a **one-vs-rest** strategy, training 45 independent binary linear probes ($\theta \in R^{4096\times 1}$) with BCEWithLogitsLoss. This design choice enables: (1) per-action interpretability, (2) mixing optimal layers for each action, (3) parallel training (8-45× speedup).

Training details: AdamW optimizer (lr=5e-4, weight_decay=1e-3), cosine annealing scheduler, early stopping (patience=10), batch size 16. Hyperparameters are optimized for small datasets ($\sim$150 examples per action). We use 70/15/15 train/val/test splits with stratification.

The binary formulation handles severe class imbalance (2.2% positive, 97.8% negative) by using AUC-ROC as the primary metric, which is robust to imbalance.

Data Generation

We generate synthetic training data using GPT-3.5 role-playing (**?**), creating conversations that demonstrate specific cognitive actions. Quality validation with GPT-4 shows 88-95% consistency across attributes (Table 1 in Chen et al. (**?**)), with 158-171 diverse topics per cognitive action category ensuring broad coverage.

Results

Probe Performance

Our binary probes achieve strong performance across all 45 cognitive actions:

- **Average AUC-ROC:** 0.78 across all probes
- **Average F1 Score:** 0.68
- **Best performers:** Suspending_Judgment (0.988 AUC), Counterfactual_Reasoning (0.984 AUC)
- **Challenging actions:** Emotion_Responding (0.778 AUC), Understanding (0.837 AUC)

These results demonstrate that cognitive actions have linearly separable representations in Gemma-3-4B's activation space, supporting the "linear representation hypothesis" (**??**).

Layer Specialization

Analysis across all 30 model layers reveals distinct layer specialization patterns. Layer 9 achieves the best average performance (AUC-ROC: 0.9481), with strong performance maintained across layers 5-24. Early layers (1-4) and late layers (25-30) show degraded performance, suggesting:

- **Early layers:** Encode surface-level linguistic features
- **Mid layers:** Capture high-level cognitive abstractions
- **Late layers:** Optimize for next-token prediction, potentially overwriting cognitive representations

Interestingly, different cognitive actions have different optimal layers. For example, *divergent_thinking* is best detected at layer 22, while *pattern_recognition* peaks at layer 9. This heterogeneity enables a multi-layer inference strategy where each cognitive action uses its best-performing layer.

Application: Therapy Transcript Analysis

We applied Cogni Map to analyze a Carl Rogers therapy session, comparing cognitive action distributions between therapist and client. Key findings:

- **Therapist-dominant actions:** Perspective_taking (2.1$\times$ client frequency), Accepting (1.8$\times$), Noticing (1.7$\times$)
- **Client-dominant actions:** Reconsidering (2.3$\times$ therapist), Emotion_receiving (1.9$\times$), Self_questioning (1.6$\times$)

This pattern aligns with Rogers' person-centered therapy model, where therapists provide empathetic understanding (noticing, accepting, perspective-taking) while clients engage in self-exploration (reconsidering, emotion_receiving). Visualizations reveal sentiment correlations: positive cognitive actions (creating, connecting) correlate with positive sentiment scores, while analytical actions (evaluating, distinguishing) show sentiment-neutral patterns.

Discussion and Future Work

Relation to prior work: While Chen et al. (**?**) demonstrated probing for user demographics in conversational AI, our work shifts focus to the model's own cognitive processes. Both leverage linear probes (**?**) and representation engineering (**?**), but target fundamentally different phenomena—external user modeling vs. internal reasoning patterns.

Limitations: (1) Probes trained on synthetic data may not fully generalize to diverse real-world text; (2) Single model (Gemma-3-4B) limits generalizability; (3) Linear probes assume independence between cognitive actions, which may not hold in practice; (4) Augmented prompting may introduce artifacts.

Future directions: (1) Extending to larger models (Llama-3-70B, GPT-4) and other architectures; (2) Training on human-annotated cognitive action data; (3) Causal intervention experiments to validate probe representations; (4) Real-time cognitive action steering for controllable generation; (5) Applications to AI safety (detecting deceptive reasoning) and education (personalized tutoring with cognitive feedback).

Broader impact: Cogni Map provides transparency into AI reasoning, potentially helping users identify biased or flawed logic, educational systems understand student thought processes, and researchers develop more interpretable AI systems. However, cognitive action detection could enable manipulation if misused (e.g., detecting vulnerable reasoning patterns for exploitation).

Conclusion

We presented Cogni Map, a system for real-time detection of 45 cognitive actions in language model activations. Through optimized activation capture and binary linear probes, we achieve 0.78 average AUC-ROC with identified layer specialization. Applications to therapy analysis demonstrate practical utility. This work bridges mechanistic interpretability and cognitive science, offering new tools for understanding how AI systems "think."

References

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016.

Leonard F. Bereska and Efstratios Gavves. Mechanistic interpretability for AI safety — a review. *arXiv preprint arXiv:2404.14082*, 2024.

Yida Chen, Aoyu Wu, Trevor DePodesta, Catherine Yeh, Kenneth Li, Nicholas Castillo Marin, Oam Patel, Jan Riecke, Shivam Raval, Olivia Seow, Martin Wattenberg, and Fernanda Viégas. Designing a dashboard for transparency and control of conversational AI. *arXiv preprint arXiv:2406.07882*, 2024.

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

Appendix A: Layer-by-Layer Performance Analysis

Figure 1 shows cognitive probe performance across all 30 layers of Gemma-3-4B. Layer 9 achieves optimal average performance (AUC-ROC: 0.9481), with a clear performance envelope between layers 5-24. This pattern suggests that cognitive abstractions are primarily encoded in middle layers, consistent with prior findings on semantic representations (**?**).

Per-Action Layer Preferences: Different cognitive actions exhibit distinct layer preferences. Metacognitive actions (meta_awareness, self_questioning) peak at layers 18-22, while creative actions (divergent_thinking, analogical_thinking) peak at layers 20-25. Analytical actions show more uniform performance across mid-layers (8-18), suggesting they are more fundamental to the model's computation.

Appendix B: Therapy Transcript Analysis Details

We analyzed a complete Carl Rogers person-centered therapy session transcript containing 1,247 therapist utterances and 1,089 client utterances. Using our best multi-layer probe configuration, we detected cognitive actions at the utterance level and computed frequency distributions.

Visualization 1 - Cognitive Action Comparison: Side-by-side bar charts reveal stark differences in cognitive action frequencies between therapist and client. The therapist's top-5 actions are: Noticing (18.3%), Perspective_taking (15.7%), Accepting (12.4%), Understanding (11.2%), Connecting (8.9%). The client's top-5 are: Reconsidering (16.8%), Emotion_receiving (14.2%), Self_questioning (11.7%), Understanding (10.3%), Analyzing (9.1%).

Visualization 2 - Cognitive Action Bias: Log_2 ratio of therapist frequency to client frequency identifies therapist-dominant patterns (positive values) vs. client-dominant patterns (negative values). The most therapist-dominant action is Perspective_taking ($\log_2$ ratio: 1.07), while the most client-dominant is Reconsidering ($\log_2$ ratio: -1.20).

Visualization 3 - Sentiment Analysis: Mean sentiment scores (from regression probes trained on layers 1-11, optimal layer 7, R^2=0.851) correlate with cognitive actions. Creative actions (Creating: +0.82, Divergent_thinking: +0.71) associate with positive sentiment, while critical actions (Evaluating: -0.12, Distinguishing: -0.08) are sentiment-neutral, and self-critical actions (Self_questioning: -0.34) associate with negative sentiment.

Clinical Insights: The detected patterns align with Rogers' theoretical framework where therapists provide "unconditional positive regard" (manifested as Accepting, Perspective_taking) while clients engage in self-directed change (Reconsidering, Self_questioning). The sentiment correlations suggest that cognitive reframing activities are accompanied by affective shifts, supporting the therapeutic mechanism of emotional processing through cognitive restructuring.

Appendix C: Implementation Details

Hardware: All experiments run on a single NVIDIA A100 GPU (80GB VRAM) with 96GB RAM. Training one linear probe takes $\sim$3 minutes. Activation capture for 6,750 examples (all layers simultaneously) takes $\sim$2.5 hours.

Software: Python 3.10, PyTorch 2.0, HuggingFace Transformers 4.35, nnsight 0.2 (third-party library for activation intervention).

Data Format: Activations stored in HDF5 format with float32 dtype (converted from bfloat16 for compatibility). Each HDF5 file contains train/val/test groups with 'activations' (N × 4096) and 'labels' (N,) datasets.

Code availability: Full source code, trained probes, and datasets available at: `github.com/ChuloIva/Cogni_map`