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

Understanding cognitive processes in large language models remains challenging, as we lack systematic tools to track internal reasoning patterns during text generation. To address this disconnect, we developed Cogni Map—a mechanistic interpretability tool leveraging linear probes to detect 45 cognitive actions spanning metacognitive, analytical, creative, and emotional categories. Binary one-vs-rest probes were trained on internal activations from 30 layers of Gemma-3-4B using 31,500 synthetic examples with augmented prompting for consistent extraction. Our approach achieved 0.78 average AUC-ROC across all cognitive actions, with layer 9 reaching peak performance (0.948 AUC). Distinct layer specialization emerged: mid-layers (5-24) capture high-level cognitive abstractions while early and late layers focus on surface features and next-token prediction. Application to therapy transcripts revealed therapist-dominant actions (*perspective_taking*, *accepting*) versus client-dominant patterns (*reconsidering*, *self_questioning*), aligning with person-centered therapy principles. Cogni Map bridges mechanistic interpretability and cognitive science, offering both quantitative probe analysis and qualitative exploration through an interactive terminal interface for understanding AI reasoning in diverse applications.

Introduction

Understanding the internal reasoning of large language models (LLMs) is crucial for ensuring their safety, interpretability, and alignment (Bereska and Gavves 2024). While previous studies have focused on profiling user attributes from conversations (Chen et al. 2024), less attention has been given to identifying the *cognitive actions* exhibited by the model during text generation.

This project introduces Cogni Map, a mechanistic interpretability tool designed to explore and annotate 45 cognitive actions spanning metacognitive, analytical, creative, and emotional categories. These actions, inspired by taxonomies from cognitive psychology, offer a detailed vocabulary for describing an AI's "thought processes," from *pattern_recognition* and *hypothesis_generation* to *emotional_reappraisal*. Cogni Map supports both quantitative analysis of cognitive patterns and qualitative exploration via an interactive TUI. The methodology is built upon linear probing techniques (Alain and Bengio 2016) to extract representations of cognitive actions from transformer activations, enabling researchers to observe the cognitive functions that are active during generation.

Contributions: (1) A synthetic dataset of 31,500 examples with 700 per cognitive action; (2) Binary probes trained across 30 layers of Gemma-3-4B achieving 0.78 average AUC-ROC with layer specialization; (3) A toolkit for quantitative and qualitative analysis; (4) Practical application to therapy transcript analysis.

Methodology

Cognitive Action Taxonomy. A taxonomy of 45 cognitive actions was defined, organized into four categories: *Metacognitive* (13 actions, e.g., *reconsidering*, *updating_beliefs*), *Analytical* (12 actions, e.g., *analyzing*, *evaluating*), *Creative* (6 actions, e.g., *divergent_thinking*, *reframing*), and *Emotional* (14 actions, e.g., *emotional_reappraisal*).

Activation Capture and Probing. Following established probing methodologies (Alain and Bengio 2016; Chen et al. 2024), activations were extracted from 30 of the 35 layers of Gemma-3-4B using nnsight. During both activation capture and inference, inputs were augmented with task-specific suffixes ("\n\nThe cognitive action being demonstrated here is" for cognitive probes; "\n\nThe sentiment of this section is" for sentiment probes) to ensure consistent extraction from the final token representation. A **one-vs-rest** strategy was used to train 45 independent binary linear probes, which allows for per-action interpretability and the flexibility to mix optimal layers for each action during inference. Training was performed with an AdamW optimizer, cosine annealing scheduler, and early stopping, using an AUC-ROC metric to handle the severe class imbalance.

Data Generation. The training dataset consists of 31,500 synthetic examples generated using Gemma 3 4b. This includes 700 examples for each of the 45 cognitive actions and 1,800 examples for sentiment analysis.

Results

Probe Performance. The binary probes demonstrated strong performance, achieving an average AUC-ROC of

0.78 and an average F1 score of 0.68 across all 45 cognitive actions. Top-performing probes included *suspending_judgment* (0.988 AUC) and *counterfactual_reasoning* (0.984 AUC), while more challenging actions included *emotion_responding* (0.778 AUC). These results confirm that cognitive actions have linearly separable representations within Gemma-3-4B's activation space (Alain and Bengio 2016).

Layer Specialization. A distinct pattern of layer specialization was observed across the 30 analyzed layers. Layer 9 yielded the best average performance (AUC-ROC: 0.9481), with a strong performance envelope across layers 5-24. Performance degraded in the earliest and latest layers, suggesting that early layers focus on surface-level features, mid-layers capture high-level cognitive abstractions, and late layers optimize for next-token prediction, potentially overwriting these representations. Notably, different cognitive actions peaked at different layers—for instance, *divergent_thinking* was best detected at layer 22, whereas *pattern_recognition* peaked at layer 9.

Application: Therapy Transcript Analysis. Cogni Map was applied to analyze a therapy session transcript, comparing the cognitive actions of the therapist and the client. The analysis revealed that therapist-dominant actions included *perspective_taking*, *accepting*, and *noticing*. In contrast, client-dominant actions were *reconsidering*, *emotion_receiving*, and *self_questioning*. These findings align with the principles of person-centered therapy (Rogers 1951), where the therapist fosters an environment for the client's self-exploration.

Discussion and Future Work

Cogni Map builds upon prior work in interpretability (Alain and Bengio 2016; Chen et al. 2024) but shifts the focus from modeling external users to tracking the internal cognitive processes of the model itself. While Chen et al. demonstrated probing for user demographics, we track cognitive actions in model-generated content—a complementary approach to representation engineering work (Zou et al. 2023).

Limitations of this study include the reliance on synthetic data, the focus on a single model (Gemma-3-4B) and the assumption of independence between cognitive actions.

Future directions include extending the tool to larger models, training on human-annotated data, and exploring applications in AI safety, ToM in LLMs and AI alignment.

Broader Impact: This tool offers a means to explore AI reasoning patterns, which could help identify flawed logic, understand student thought processes in educational settings, and build more interpretable AI systems. The combination of quantitative probes and a qualitative TUI supports a wide range of use cases.

Conclusion

Cogni Map is a practical tool for exploring and annotating 45 cognitive actions in language models. By using linear probes on the internal activations of Gemma-3-4B, this work successfully identified and analyzed these actions, revealing specialized layer-dependent representations. The toolkit, which supports both quantitative and qualitative analysis, has demonstrated its utility in a real-world application, bridging the gap between mechanistic interpretability and cognitive science. The full source code, trained probes, and datasets are available on GitHub.

References

Alain, G.; and Bengio, Y. 2016. Understanding intermediate layers using linear classifier probes. arXiv:1610.01644.

Bereska, L.; and Gavves, E. 2024. Mechanistic Interpretability for AI Safety — A Review. arXiv:2404.14082.

Chen, Y.; Wu, A.; DePodesta, T.; Yeh, C.; Li, K.; Castillo Marin, N.; Patel, O.; Riecke, J.; Raval, S.; Seow, O.; Wattenberg, M.; and Viégas, F. 2024. Designing a Dashboard for Transparency and Control of Conversational AI. arXiv:2406.07882.

Rogers, C. R. 1951. *Client-Centered Therapy: Its Current Practice, Implications, and Theory*. Boston: Houghton Mifflin.

Zou, A.; Phan, L.; Chen, S.; Campbell, J.; Guo, P.; Ren, R.; Pan, A.; Yin, X.; Mazeika, M.; Dombrowski, A.-K.; Goel, S.; Li, N.; Byun, M. J.; Wang, Z.; Mallen, A.; Basart, S.; Koyejo, S.; Song, D.; Fredrikson, M.; Kolter, J. Z.; and Hendrycks, D. 2023. Representation Engineering: A Top-Down Approach to AI Transparency. arXiv:2310.01405.

Additional Visualizations

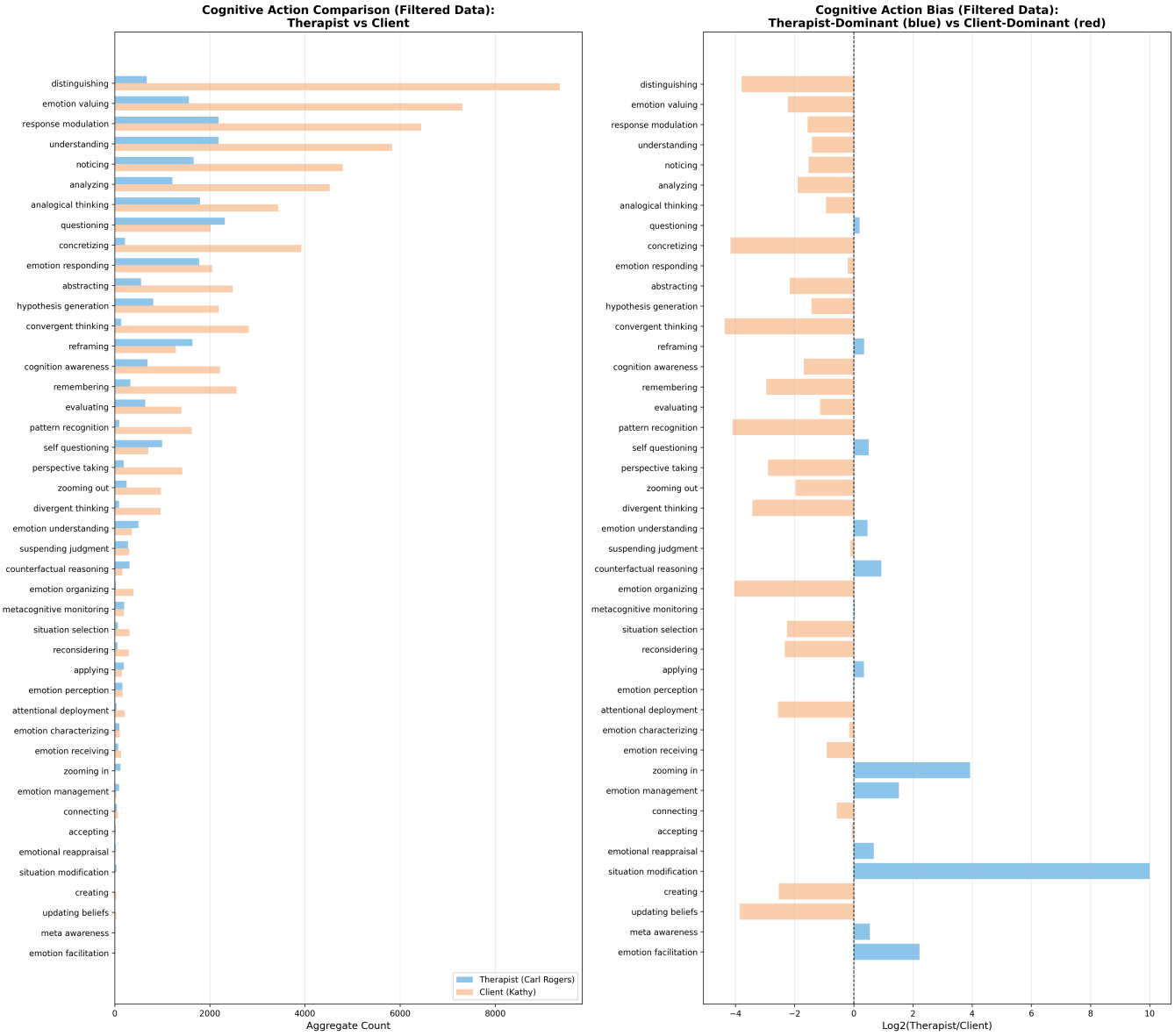

Figure 1: Therapy session analysis comparing cognitive action patterns between therapist (Carl Rogers) and client (Kathy). Left panel shows aggregate counts of cognitive actions for both participants across filtered data. Right panel displays cognitive action bias as log2 ratio, highlighting therapist-dominant actions (blue) such as *situation_modification*, *emotional_reappraisal*, and *emotion_facilitation*, versus client-dominant actions (red) including *zooming_in*, *self_questioning*, and *applying*. These patterns align with person-centered therapy principles.

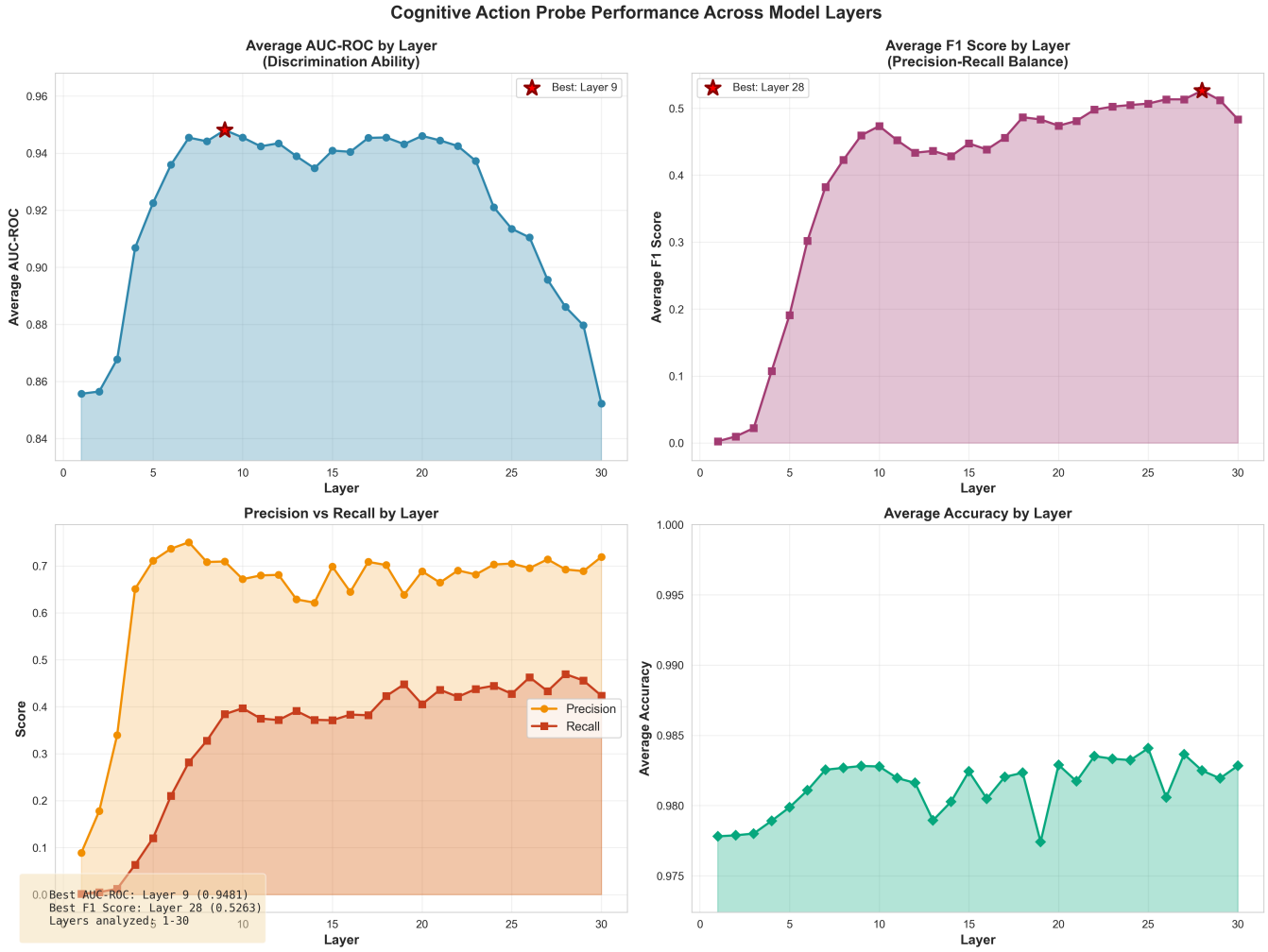

Figure 2: Cognitive action probe performance across all 30 layers of Gemma-3-4B. The visualization shows average AUC-ROC scores for each layer, with Layer 9 achieving peak performance (0.948 AUC-ROC). Strong performance is maintained across mid-layers (5-24), while early and late layers show degraded performance. This pattern suggests that early layers focus on surface-level features, mid-layers capture high-level cognitive abstractions, and late layers optimize for next-token prediction, potentially overwriting intermediate representations.

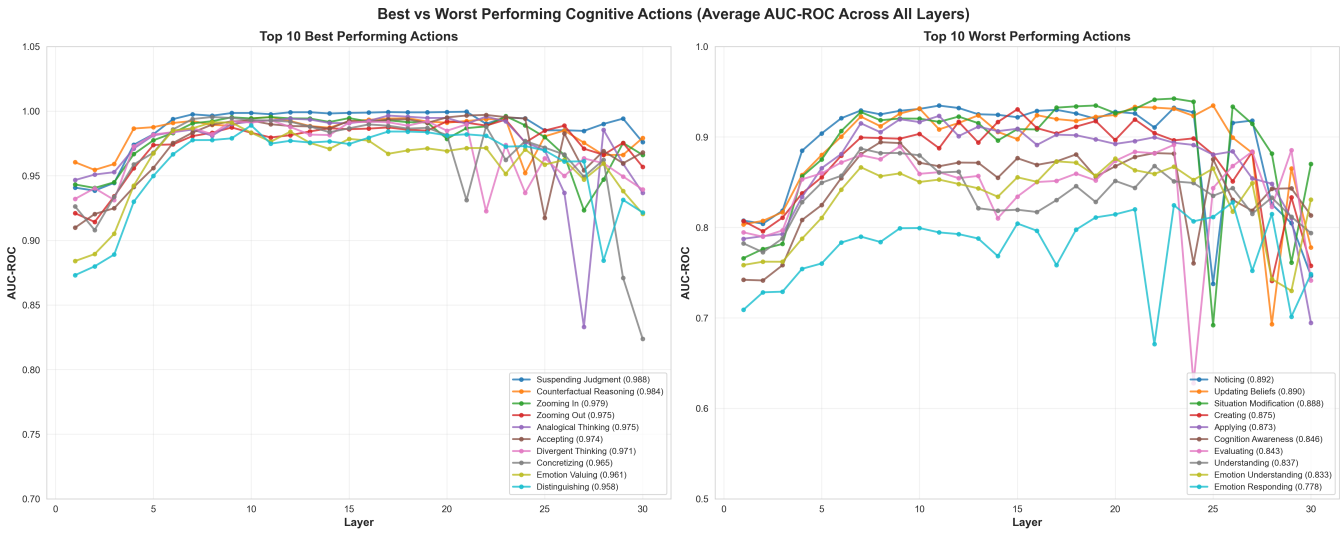

Figure 3: Comparison of top 10 and bottom 10 performing cognitive actions ranked by average AUC-ROC across all layers. Best performers like *suspending_judgment* (0.988) and *counterfactual_reasoning* (0.984) show consistently high performance and distinct activation patterns across most layers. Worst performers like *emotion_responding* (0.778) and *understanding* (0.837) exhibit more variability and lower overall discrimination ability, suggesting these concepts may be more distributed or context-dependent in the model's representation space.