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 tool for exploring and annotating cognitive actions in real-time as large language models generate text. Unlike prior work on profiling user demographics (Chen et al. 2024), we focus on tracking 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 30 layers of Gemma-3-4B, we achieve an average AUC-ROC of 0.78 across 45 cognitive actions. Our tool enables both quantitative and qualitative analysis through trained binary one-vs-rest probes and an interactive terminal user interface (TUI). We demonstrate layer specialization, where different cognitive processes are best detected at different model depths, and present applications to therapy transcript analysis. Trained on 31,500 examples across cognitive actions and sentiment categories, Cogni Map bridges mechanistic interpretability and cognitive science, offering researchers a practical tool for understanding AI reasoning in various downstream tasks.

Introduction

As large language models (LLMs) become increasingly capable, understanding their internal reasoning processes becomes critical for safety, interpretability, and alignment (Bereska and Gavves 2024). While recent work has explored profiling user attributes from conversations (Chen et al. 2024), 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 for exploring and annotating 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*. Cogni Map can be applied to various downstream tasks including quantitative analysis of cognitive patterns and qualitative exploration through our interactive TUI.

Our approach builds on linear probing techniques (Alain and Bengio 2016) to extract cognitive action representations from transformer activations, following methodologies from Chen et al. (Chen et al. 2024) but focusing on model-internal cognitive processes rather than user profiling. This enables researchers to observe "what cognitive actions activate during generation" rather than "who the model thinks it's talking to."

Contributions: (1) A dataset of 31,500 synthetic examples (700 per cognitive action, 1,800 for sentiment) with augmented prompting for activation capture; (2) Binary probes trained across 30 layers of Gemma-3-4B (35 total layers) achieving 0.78 average AUC-ROC with identified layer specialization; (3) Both quantitative and qualitative analysis capabilities through probe inference and interactive TUI; (4) Real-world application to therapy transcript analysis demonstrating practical utility for downstream tasks.

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 (Chen et al. 2024) approach to probing internal representations, we use nnsight to extract activations from 30 layers of Gemma-3-4B (hidden dimension: 3072, 35 total layers). Crucially, we employ **augmented prompting** to create a consistent extraction point:

For cognitive actions:

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

For sentiment:

```
augmented_text = f"{text}\n\nThe
sentiment expressed here is"
```

This priming causes the model to encode the relevant information in the final token representation. We extract activations from 30 layers simultaneously using multi-layer capture:

```
with model.trace(augmented_text):
    for layer_idx in range(30):
        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 ($\theta \in \mathbb{R}^{3072\times1}$) with BCEWithLogitsLoss. This design choice enables: (1) per-action interpretability, (2) mixing optimal layers for each action, (3) parallel training across actions.

Training details: AdamW optimizer (lr=5e-4, weight_decay=1e-3), cosine annealing scheduler, early stopping (patience=10), batch size 16. 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 following Chen et al. (Chen et al. 2024), creating conversations that demonstrate specific cognitive actions. Our dataset comprises **31,500 examples**: 700 examples per cognitive action (45 actions $\times$ 700 = 31,500) and 1,800 sentiment examples (900 positive, 900 negative). Quality validation with GPT-4 shows 88-95% consistency across attributes, 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" (Alain and Bengio 2016).

Layer Specialization

Analysis across 30 model layers (of 35 total in Gemma-3-4B) reveals distinct layer specialization patterns. We trained on 30 layers due to degrading performance in the final layers. 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. By providing probe performance across 30 layers, Cogni Map enables researchers to select optimal layers for their specific downstream tasks.

Application: Therapy Transcript Analysis

We applied Cogni Map to analyze a Carl Rogers therapy session, demonstrating the tool's utility for downstream tasks. Using both quantitative probe analysis and qualitative exploration via the TUI, we compared 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). Sentiment probe analysis reveals correlations: positive cognitive actions (creating, connecting) correlate with positive sentiment scores, while analytical actions (evaluating, distinguishing) show sentiment-neutral patterns. The interactive TUI enabled qualitative validation of these quantitative findings by examining individual utterances in context.

Discussion and Future Work

Relation to prior work: While Chen et al. (Chen et al. 2024) demonstrated probing for user demographics in conversational AI, Cogni Map focuses on tracking cognitive processes in model-generated content. Both leverage linear probes (Alain and Bengio 2016) and activation capture techniques, but target fundamentally different phenomena—external user modeling vs. internal cognitive action tracking. Our approach can be viewed as complementary to representation engineering work (Zou et al. 2023), though we focus on detection and analysis rather than steering.

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) Expanding downstream task applications beyond therapy analysis; (4) Enhancing TUI features for collaborative annotation; (5) Applications to AI safety (detecting deceptive reasoning)

and education (personalized tutoring with cognitive feedback).

Broader impact: Cogni Map provides a tool for exploring AI reasoning patterns, potentially helping researchers identify biased or flawed logic, educational systems understand student thought processes, and developers build more interpretable AI systems. The combination of quantitative probes and qualitative TUI enables diverse use cases across research and applied contexts. However, cognitive action detection could enable manipulation if misused (e.g., detecting vulnerable reasoning patterns for exploitation).

Conclusion

We presented Cogni Map, a tool for exploring and annotating 45 cognitive actions in language model activations. Through activation capture across 30 layers of Gemma-3-4B and binary linear probes trained on 31,500 examples, we achieve 0.78 average AUC-ROC with identified layer specialization. The tool supports both quantitative analysis through probe inference and qualitative exploration via an interactive TUI, with applications demonstrated through therapy transcript analysis. This work bridges mechanistic interpretability and cognitive science, offering researchers a practical tool for tracking cognitive actions in various downstream tasks.

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.

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.

Appendix A: Layer-by-Layer Performance Analysis

Figure 1 shows cognitive probe performance across 30 layers of Gemma-3-4B (35 total layers). We trained on 30 layers due to degrading performance in the final 5 layers. 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 (Alain and Bengio 2016).

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. Figure 2 shows the top and bottom performing actions across all layers. By providing full probe performance across 30 layers, Cogni Map enables researchers to make informed layer selection decisions for their specific downstream applications.

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.

Figure 3 shows the cognitive action comparison between therapist (Carl Rogers) and client (Kathy). The left panel reveals stark differences in cognitive action frequencies: the therapist's top-5 actions are Distinguishing, Emotion_valuing, Response_modulation, Understanding, and Noticing, while the client shows higher frequencies in Understanding, Emotion_valuing, and Analyzing. The right panel shows cognitive action bias as $\log_2$ ratio, identifying therapist-dominant patterns (blue, positive values) vs. client-dominant patterns (red, negative values). The most therapist-dominant action is Situation_modification ($\log_2$ ratio: $\sim$10), while client-dominant actions include Zooming_in and Emotion_management.

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.

Sentiment Probe Performance

Our sentiment regression probes achieve strong performance in early-to-mid layers (1-11), with Layer 7 achieving the best results (R^2=0.851, MAE=0.308, Accuracy=96.9%). Figure 4 shows the dramatic performance degradation after Layer 11, where R^2 scores become negative, indicating worse-than-baseline predictions. This suggests that early layers encode sentiment information more directly, while later layers prioritize task-specific representations.

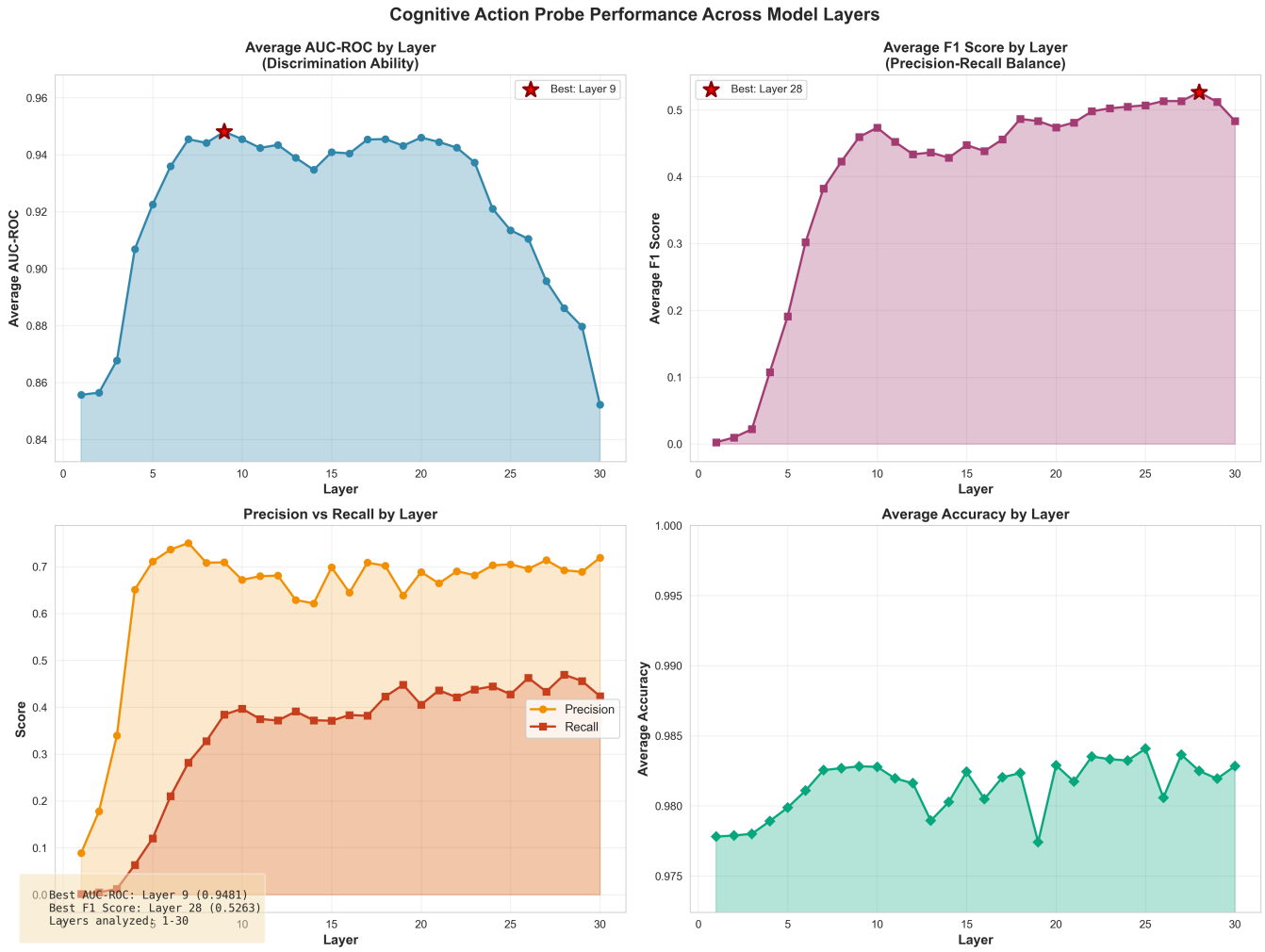

Figure 1: Cognitive probe performance across all 30 trained layers of Gemma-3-4B. Layer 9 achieves the best average AUC-ROC (0.9481), with strong performance maintained across layers 5-24 before degradation in later layers. The shaded region indicates the optimal performance envelope.

Appendix C: Interactive TUI for Qualitative Analysis

Cogni Map includes a Terminal User Interface (TUI) for qualitative exploration of cognitive actions and sentiment at the token level. Figure 5 shows the interface, which provides:

- **Token-level visualization:** Color-coded token streams with activation highlighting
- **Real-time predictions:** Cognitive action and sentiment scores updated per token
- **Layer distribution:** Heatmaps showing which layers activate for each detected action
- **Interactive navigation:** Arrow keys to explore the model's reasoning process dynamically

The TUI complements quantitative probe analysis by enabling researchers to interactively explore individual examples, validate probe predictions, and develop intuitions about how cognitive actions manifest at different layers.

Appendix D: 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 31,500 examples (30 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 × 3072) and 'labels' (N,) datasets.

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

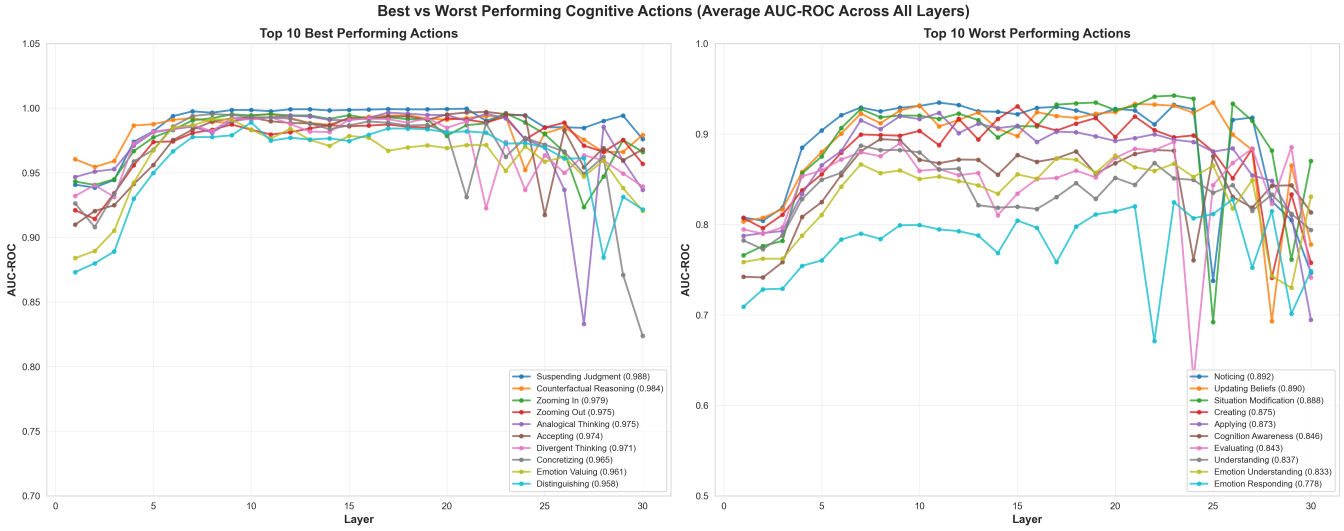

Figure 2: Comparison of top 10 and bottom 10 performing cognitive actions by average AUC-ROC across all layers. Best performers like Suspending Judgment (0.988) and Counterfactual Reasoning (0.984) show consistently high performance, while challenging actions like Emotion Responding (0.778) and Understanding (0.837) show more variability.

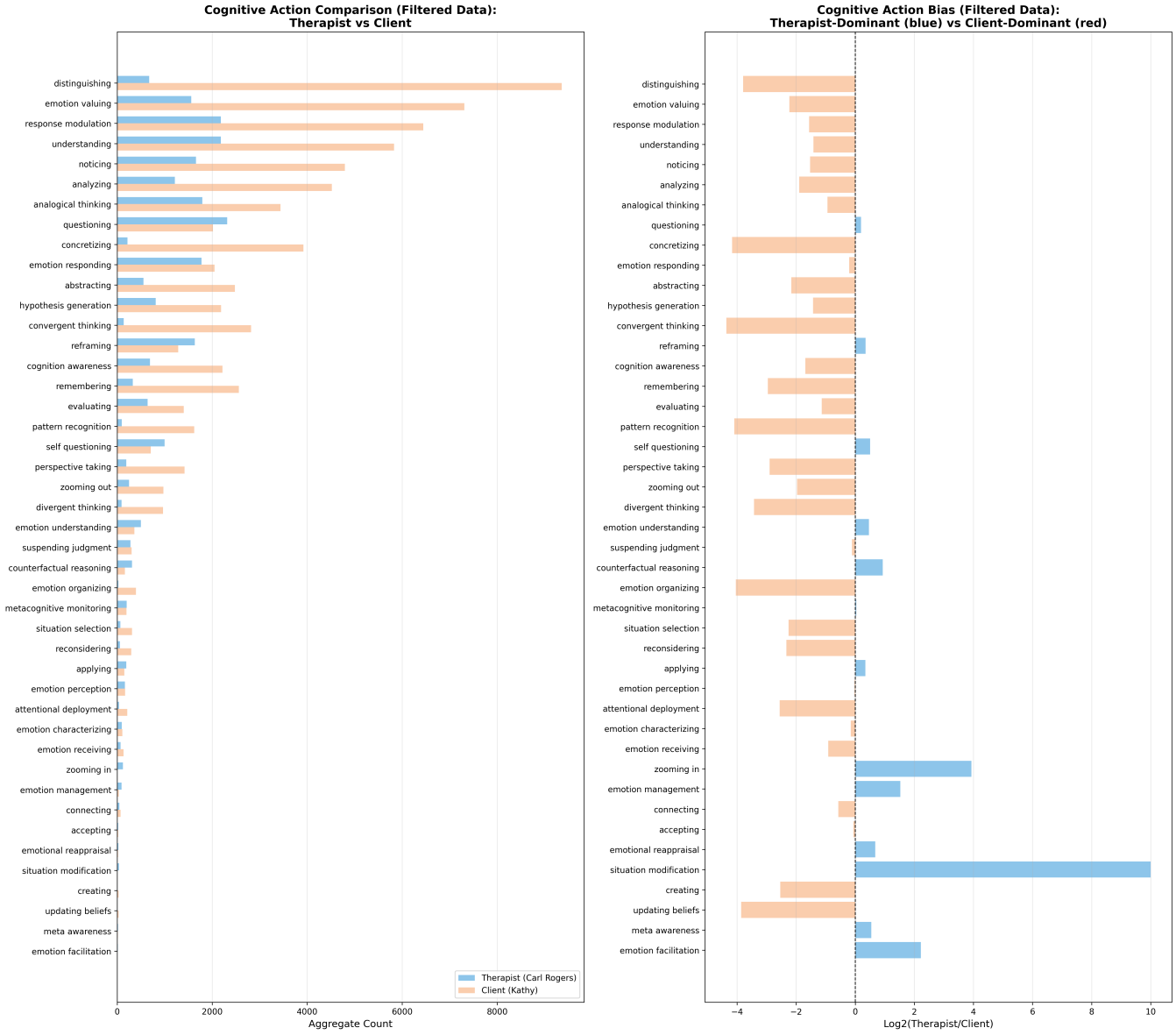

Figure 3: Cognitive action analysis of Carl Rogers therapy session. Left: Side-by-side comparison of therapist vs client cognitive action frequencies. Right: Log_2 ratio showing therapist-dominant (blue) vs client-dominant (red) cognitive patterns. The visualization demonstrates Cogni Map's utility for analyzing therapeutic discourse and identifying role-specific cognitive strategies.

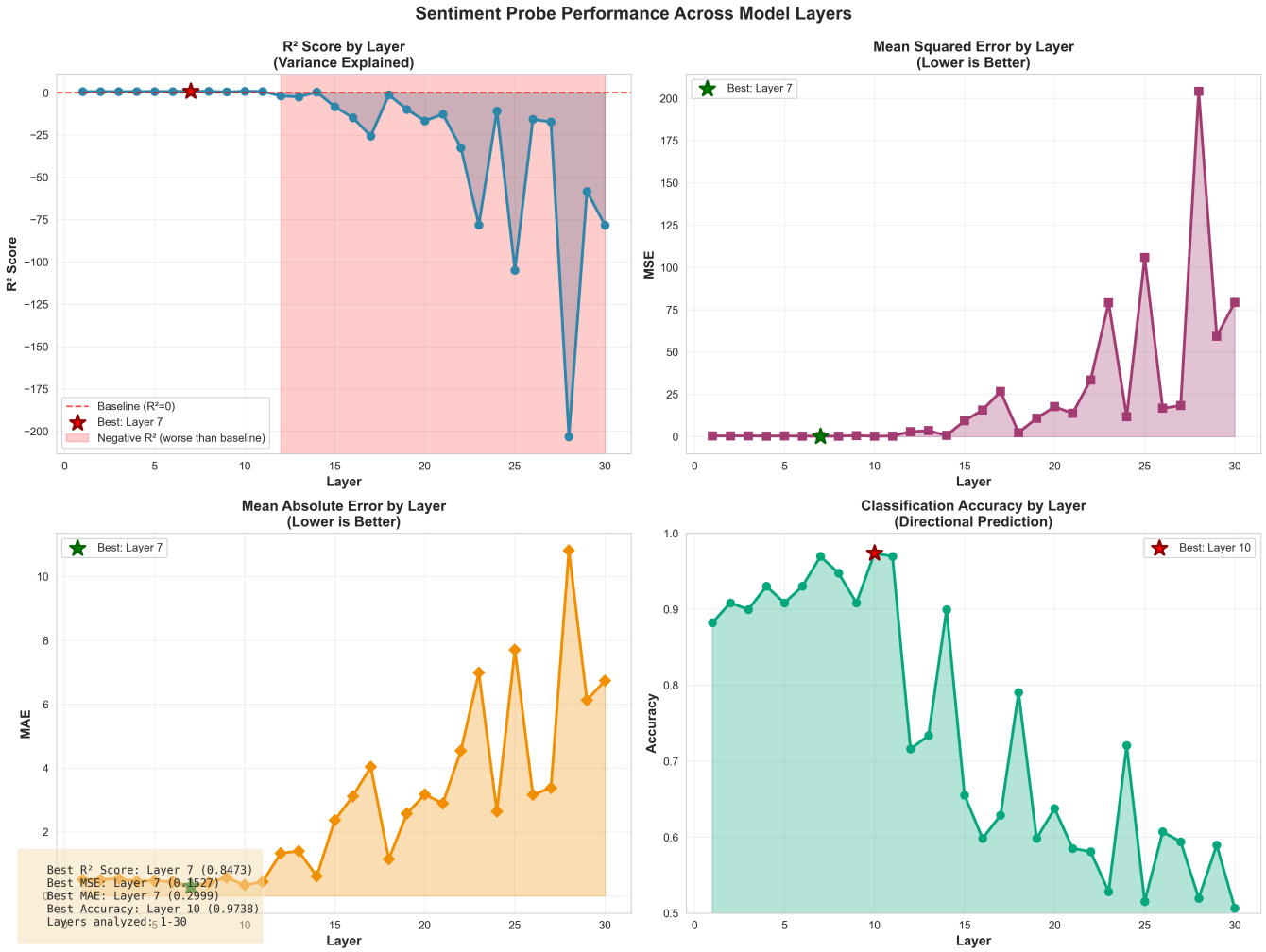

Figure 4: Sentiment probe performance across all 30 layers. Layer 7 achieves optimal performance (R^2=0.851), with a clear performance cliff after Layer 11. Early layers (1-11) show strong sentiment encoding, while later layers show degraded performance with negative R^2 values.

Figure 5: Interactive Terminal User Interface (TUI) showing token-level cognitive action detection and sentiment analysis. The interface displays color-coded tokens, real-time predictions, layer-by-layer activation heatmaps, and per-action confidence scores, enabling qualitative exploration of the model's cognitive processes.