Brije: A General Framework for Real-Time Detection of Cognitive Constructs in Language Models

Anonymous submission

Abstract

This paper introduces Brije, a general framework for real-time detection and analysis of cognitive constructs in language models. While we demonstrate the system with 45 cognitive actions—such as analyzing, reconsidering, and perspective-taking—the framework is designed to detect *any* cognitive construct for which sufficient quality training data can be generated. Brije packages a complete pipeline: (1) an LLM-based synthetic data generator using scientifically-grounded taxonomies, (2) a probe training system that learns to detect constructs from model activations, and (3) a real-time streaming inference engine with interactive visualization. We validate the approach by analyzing 133 therapy session transcripts, revealing cognitive strategies employed by therapists and clients. Our work provides a generalizable method for interpreting internal reasoning processes in language models, with applications in AI safety, transparency, and human-AI interaction across diverse domains.

Introduction

Large language models have achieved remarkable performance across diverse tasks, yet their internal reasoning processes remain opaque. Understanding *how* these models arrive at their outputs—not merely *what* they produce—is critical for ensuring safety, alignment, and effective deployment in high-stakes domains such as mental health, education, and decision support.

The Challenge. Current evaluation methods focus primarily on final outputs, treating LLMs as black boxes. This approach provides limited insight into the cognitive strategies models employ during generation, making it difficult to detect harmful reasoning patterns, validate alignment, or understand failure modes.

Our Solution. We introduce Brije, a general-purpose framework for detecting cognitive constructs in language models at the token level—analogous to functional MRI for neural activity. While traditional mechanistic interpretability research focuses on hand-crafted interpretations of specific neurons or circuits, Brije provides a *scalable methodology*: define any cognitive construct, generate training data, and train probes to detect it in real-time. The framework packages three integrated components:

1. Data Generation Pipeline. An LLM-based synthetic data generator that creates diverse, high-quality training examples for any cognitive construct. Using stratified sampling, parallel async processing, and scientifically-grounded taxonomies, the system can generate thousands of labeled examples with controlled variation in domain, complexity, style, and context.

2. Probe Training System. A novel augmented prompting technique for extracting discriminative activations from model internals, combined with binary classification probes trained using one-vs-rest methodology. The system automatically handles data preparation, training, validation, and multi-layer probe ensembles.

3. Real-Time Inference Engine. A streaming token-by-token analysis system with interactive terminal visualization, configurable thresholds, layer-wise analysis, and export capabilities for statistical analysis.

Key Contributions:

- A general framework for detecting arbitrary cognitive constructs in LLMs, not limited to predefined taxonomies
- Complete open-source pipeline from data generation through probe training to real-time inference
- Demonstration with 45 cognitive actions spanning reasoning, metacognition, and emotional processing
- Novel augmented prompting technique for training discriminative probes on internal model states
- Analysis of 133 therapy transcripts (1,935 utterances) validating the approach in a real-world domain
- All code, data generation tools, pre-trained probes, and analysis tools released for community use

Related Work

Mechanistic Interpretability. Recent work has systematically reviewed interpretability methods for transformer-based models (Rai, Zhou et al. 2025), including probing, activation patching, and sparse autoencoders. The ICML 2024 Mechanistic Interpretability Workshop showcased 93 papers advancing this field. Our work extends these approaches by focusing on real-time, multi-action detection at the token level.

Probing Classifiers. Probing has emerged as a diagnostic technique for analyzing neural network representations (Belinkov 2022). Edge probing frameworks have compared contextualized representations across models like BERT and

GPT (Tenney, Das, and Pavlick 2019). However, these approaches typically focus on linguistic features rather than cognitive processes, and lack real-time streaming capabilities.

Internal State Analysis. Recent work has demonstrated that LLMs can monitor and control their internal activations (Authors 2025), revealing a ”metacognitive space” of lower dimensionality than the full neural space. Methods like INSIDE (Authors 2024a) exploit internal states for hallucination detection. Brije extends this line of work to a comprehensive cognitive action taxonomy.

AI for Mental Health. LLM-based conversational systems show promise for therapeutic applications (Authors 2024b,c), but evaluation remains non-standardized. Brije provides a novel lens for analyzing therapeutic AI by detecting cognitive actions in real-time conversations.

Methodology

Framework Overview

Brije is designed as a *general tool* for detecting cognitive constructs in language models. The key insight is that *any* construct can be detected if we can generate sufficient quality training data. We demonstrate this generality through a case study with 45 cognitive actions, but the framework applies to any taxonomy researchers wish to study—from deceptive reasoning patterns to theory-of-mind capabilities to domain-specific expertise markers.

Model Configuration

We use Gemma 3 4B (instruction-tuned) as our base model, analyzing layers 21-30 (75-95% depth) where high-level reasoning emerges. The framework is model-agnostic and can be applied to any transformer-based LLM with accessible internal activations.

Demonstration: 45 Cognitive Actions

To validate the framework, we implement detection for 45 cognitive actions organized into five scientifically-grounded categories:

Core Reasoning (19 actions): reconsidering, reframing, noticing, perspective-taking, questioning, abstracting, connecting, distinguishing, updating beliefs, pattern recognition, analogical thinking, counterfactual reasoning, hypothesis generation, meta-awareness, and others.

Bloom’s Taxonomy (6 actions): remembering, understanding, applying, analyzing, evaluating, creating.

Guilford’s Operations (3 actions): divergent thinking, convergent thinking, cognition awareness.

Metacognitive (3 actions): metacognitive monitoring, metacognitive regulation, self-questioning.

Affective Regulation (14 actions): emotional reappraisal, emotion receiving, responding, valuing, organizing, characterizing, situation selection/modification, attentional deployment, response modulation, emotion perception, facilitation, understanding, management.

However, the framework is not limited to these actions. Researchers can define their own constructs and use the data generation pipeline to create training data for detection.

Component 1: Data Generation Pipeline

The data generation system is a core contribution that enables Brije’s generality. Rather than manually collecting and labeling training data, we use an LLM-based synthetic generation approach that can scale to any cognitive construct.

Architecture. The pipeline consists of three modules packaged in `third_party/datagen/`:

- `variable_pools.py`: Defines construct taxonomies and contextual variables (domains, subjects, emotional states, perspectives, complexity levels)
- `prompt_templates.py`: Four template types for format diversity (single-action 70%, action-chains 20%, dialogue 5%, thought-stream 5%)
- `data_generator.py`: Async parallel generation engine with stratified sampling, automatic checkpointing, and error handling

Generation Process. For each cognitive action, the system:

1. Samples contextual variables (domain, subject, emotional state, complexity)
2. Selects a template type according to the distribution
3. Constructs a prompt with variable substitution
4. Generates text via Ollama API (gemma3:27b) with async parallelism
5. Validates output quality and saves with metadata

Performance. Using 16 parallel requests on a 40GB GPU, the system generates $\sim$7,000 examples in 3.7 hours with automatic checkpointing every 100 examples. Generated data achieves high diversity (36 domains, 50+ subjects each) and stratification (155 examples per action ±5).

Generalization. To detect new constructs, researchers simply: (1) define the construct and examples in `variable_pools.py`, (2) optionally add custom templates, and (3) run the generator. No manual labeling required.

Component 2: Probe Training System

Augmented Activation Extraction. We introduce a novel technique for capturing discriminative activations. For each training example (e.g., ”I was reconsidering my approach...”), we append the prompt ”\n\nThe cognitive action being demonstrated here is” and extract the last token’s activation. This priming causes the model to ”think about” cognitive actions, producing more separable representations in activation space.

Binary Classification. We train one binary classifier per construct using a one-vs-rest approach. Each probe is a linear layer with optional multi-head attention, trained with binary cross-entropy loss, AdamW optimization (lr=5e-4), and early stopping. For our 45-action demonstration, this yields 450 total probes (45 actions × 10 layers).

Performance. Across all 45-action probes, we achieve an average AUC-ROC of 0.78 and F1-score of 0.68. Best-performing actions (noticing, analyzing, hypothesis generation) exceed 0.85 AUC, while fine-grained emotional distinctions remain more challenging (0.65-0.70 AUC). Performance varies by construct complexity—coarse distinctions achieve higher accuracy than fine-grained ones.

Component 3: Real-Time Streaming Inference

The streaming engine enables token-by-token analysis during text generation. For each token position t, we:

1. Decode tokens $[0 : t+1]$ and apply augmented prompting
2. Extract activations from target layers (e.g., 21-30 for Gemma 3 4B)
3. Apply all trained probes to obtain confidence scores
4. Aggregate predictions across layers and visualize

The interactive terminal UI (implemented in `src/probes/Interactive_TUI.py`) displays active cognitive constructs with confidence scores, layer distributions, and peak activation tokens. Users can adjust detection thresholds, filter by layer, and export detailed CSV files for statistical analysis. The system supports real-time streaming during model inference or batch processing of existing transcripts.

Validation: Therapy Dataset Analysis

To validate that Brije's detections correspond to meaningful cognitive patterns in real-world data, we applied the framework to the AnnoMI corpus containing 133 motivational interviewing therapy transcripts. We focused on analyzing Carl Rogers' therapy sessions, comprising 1,935 utterances (860 therapist, 1,075 client). This demonstrates the framework's practical utility when applied to domain-specific analysis tasks.

Therapist Cognitive Patterns. The most frequent therapist actions were: (1) response modulation—managing emotional expression to create safety, (2) noticing—detecting client states and patterns, (3) hypothesis generation—inferring underlying issues, and (4) emotion perception—recognizing emotional content. Network analysis revealed ”noticing” as the central hub, co-occurring with most other actions.

Client Cognitive Patterns. Clients exhibited high frequencies of self-questioning, reconsidering beliefs, and emotional reappraisal—indicating active cognitive restructuring during therapy. Moments of perspective-taking often preceded breakthroughs in understanding.

Layer Specialization. Early layers (21-23) activated primarily for basic pattern recognition and noticing. Middle layers (24-27) showed strong signals for perspective-taking and distinguishing. Late layers (28-30) specialized in metacognitive monitoring and hypothesis generation—suggesting hierarchical cognitive processing.

Temporal Dynamics. Token-level analysis revealed that cognitive actions often activate in sequences: noticing $\rightarrow$ hypothesis generation $\rightarrow$ perspective-taking, mirroring therapeutic reasoning progressions.

Discussion

A Tool, Not Just a Taxonomy. While we demonstrate Brije with 45 cognitive actions, the central contribution is the *framework itself*. Researchers can adapt the system to detect any cognitive constructs of interest—theory-of-mind reasoning, deceptive patterns, creativity markers, domain expertise indicators, or alignment-relevant cognitive strategies. The complete pipeline (data generation, probe training, real-time inference) is packaged and released open-source in `third_party/datagen/` and the main repository.

Scalability and Generalization. The data generation pipeline eliminates the bottleneck of manual labeling. Defining new constructs requires only: (1) specifying the construct definition and examples in `variable_pools.py`, (2) running the async generator for a few hours, and (3) training probes on the resulting data. This makes Brije practical for exploring diverse research questions without extensive human annotation.

Implications for AI Safety. The framework enables detection of cognitive patterns associated with deception, manipulation, or reasoning failures. Real-time monitoring could flag concerning cognitive strategies before they manifest in outputs, supporting more robust alignment verification. Researchers can define safety-relevant constructs and generate training data to detect them.

Domain-Specific Applications. Beyond therapy analysis, Brije can be adapted to domain-specific cognitive constructs—legal reasoning patterns in law, diagnostic thinking in medicine, pedagogical strategies in education. The framework's generality makes it applicable wherever researchers can define and exemplify the constructs they wish to detect.

Human-AI Interaction. Token-level cognitive transparency helps users understand AI reasoning processes, building trust and enabling more effective collaboration in sensitive domains.

Limitations. Detected ”cognitive constructs” represent statistical patterns in activation space and may not correspond precisely to human cognition or model ”intentions.” Probe accuracy varies by construct complexity, and the taxonomy reflects design choices rather than ground truth about model internals. The quality of detections depends critically on the quality of generated training data. Future work should validate patterns through causal interventions, expand coverage to other models, and develop better data quality metrics.

Conclusion and Future Work

Brije provides a general framework for detecting arbitrary cognitive constructs in language models through a complete pipeline: synthetic data generation, probe training, and real-time inference. While we demonstrate the approach with 45 cognitive actions and validate it on therapy transcripts, the framework's key value is its *generalizability*—researchers can define their own constructs and use the packaged tools to detect them.

The three-component architecture (data generator in `third_party/datagen/`, probe training system, and streaming inference engine) makes Brije practical for diverse research applications without requiring extensive manual annotation. By packaging the entire pipeline open-source, we enable the research community to explore cognitive constructs relevant to their specific domains—from AI safety to human-AI interaction to domain-specific cognitive modeling.

Future directions include: (1) expanding to other LLMs and modalities, (2) causal validation through activation steering, (3) automated data quality assessment for generated training sets, (4) detecting cognitive construct transitions and sequences, (5) applications to AI safety evaluation with alignment-relevant constructs, and (6) integration with human cognitive modeling research.

All code, the complete data generation pipeline (`third_party/datagen/`), pre-trained probes for 45 cognitive actions, and analysis tools are available at: [URL will be provided upon acceptance].

Acknowledgments

We thank the developers of nnsight for activation extraction tools and the AnnoMI team for providing therapy transcripts.

References

Authors. 2024a. INSIDE: LLMs' Internal States Retain the Power of Hallucination Detection. In *International Conference on Learning Representations*.

Authors. 2024b. Introducing CounseLLMe: A Dataset of Simulated Mental Health Dialogues for Comparing LLMs against Humans. *ScienceDirect*. Multilingual dataset with GPT-3.5, Claude-3 Haiku, and LLaMAntino.

Authors. 2024c. LLM-based Conversational AI Therapist for Daily Functioning Screening and Psychotherapeutic Intervention via Everyday Smart Devices. *ACM Transactions on Computing for Healthcare*. ArXiv:2403.10779.

Authors. 2025. Language Models Are Capable of Metacognitive Monitoring and Control of Their Internal Activations. arXiv:2505.13763.

Belinkov, Y. 2022. Probing Classifiers: Promises, Shortcomings, and Advances. *Computational Linguistics*, 48(1): 207–219.

Rai, Z.; Zhou, Y.; et al. 2025. A Practical Review of Mechanistic Interpretability for Transformer-Based Language Models. *arXiv preprint arXiv:2406.xxxxx*. Tutorial presented at ICML 2025.

Tenney, I.; Das, D.; and Pavlick, E. 2019. BERT Rediscovers the Classical NLP Pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 4593–4601.