Probing Circuit Robustness: How Syntactic Form Shapes Neural Circuit Activation in LLMs

Final Report

Arul Murugan
UC Berkeley
arul@berkeley.edu

Abstract

We investigate whether neural circuits in language models remain stable under semantic-preserving paraphrases. Using the Gemma-2-2B base model with Gemma Scope sparse autoencoders, we extract attribution graphs for 500 prompt pairs across five categories and measure circuit stability via weighted Jaccard similarity, Spearman correlation, and feature overlap. Contrary to our initial hypothesis that harm-related circuits would be less stable than factual circuits, we found harm-topic circuits exhibit *higher* stability (0.543 vs. 0.362). However, our central finding is that *syntactic form dominates circuit activation*: prompts with identical meaning but different syntactic structures (e.g., WH-questions vs. fill-in-the-blank) activate fundamentally different circuits (Jaccard as low as 0.08), while syntactically similar paraphrases maintain high overlap (Jaccard$>$0.80). This syntactic effect explains the apparent stability difference between categories. A controlled grammatical role analysis confirms this with large effect size (Cohen's d=2.45, $p<10^{-6}$). We also observe rank-overlap divergence: despite low circuit overlap, feature importance rankings remain highly correlated (mean Spearman=0.90), suggesting consistent computational priorities across different circuit instantiations. Adversarially-framed prompts show 48% lower stability than direct requests, indicating that framing substantially alters internal representations.

1 Motivation and Research Question

Large language models (LLMs) increasingly power high-stakes applications, yet their internal decision-making processes remain poorly understood. Recent work shows that models remain susceptible to adversarial attacks, as evidenced by the prevalence of jailbreak attacks on models like ChatGPT (Zou et al., 2023). A critical gap exists in understanding how models internally represent different types of content and whether these representations are robust to surface-form variations.

This project investigates a fundamental question: **Do internal circuits activated by semantically equivalent prompts remain stable when syntactic form varies?** We compare circuit stability across different content categories in the Gemma-2-2B base model, examining how harm-related prompts, factual queries, and adversarially-framed inputs activate different neural pathways. Understanding these representational patterns provides foundational insights for future work on alignment and robustness.

Unlike prior work that focuses on output-level robustness, we examine internal mechanisms directly by extracting and comparing attribution graphs (structured representations of feature interactions that produce model outputs). This approach lets us determine whether observed output brittleness stems from fundamental architectural limitations or superficial alignment failures.

2 Proposed Methodology

We use recent advances in mechanistic interpretability to conduct a controlled comparison of circuit stability. Mechanistic interpretability aims to identify structures, circuits, or algorithms encoded in model weights through causal analysis of internal components.

Model and Tools: We use Gemma-2-2B with pre-trained sparse autoencoders from Gemma Scope (Lieberum et al., 2024), which decompose model activations into interpretable features. We apply the circuit-tracer library (Hanna et al., 2025; Lindsey et al., 2025), which computes direct causal effects between transcoder features using attribution patching (Syed et al., 2024). This approach identifies interpretable feature circuits by analyzing patterns in circuit activations, revealing components responsible for different model capabilities.

Model Selection Justification: Our choice of Gemma-2-2B is driven by several methodological and practical considerations:

(1) Sparse Autoencoder Availability: Gemma Scope (Lieberum et al., 2024) provides the most comprehensive open-source suite of sparse autoencoders for mechanistic interpretability research, comprising over 400 SAEs with more than 30 million learned features trained on 4–16 billion tokens. The Gemma Scope SAEs are trained on Gemma 2 pre-trained model activations, ensuring proper alignment between our circuit extraction and the underlying model representations.

(2) Neuronpedia Infrastructure: Our circuit extraction relies on Neuronpedia (Lin, 2024), an open-source interpretability platform that provides attribution graph generation using Gemma Scope SAEs. Neuronpedia integrates with the circuit-tracer library (Hanna et al., 2025), which was developed by Anthropic researchers and open-sourced to enable attribution graph generation on popular open-weights models. This service is optimized for and validated on the base Gemma-2 models, with the Neuronpedia team reporting nearly 5,000 user-generated graphs. The coherent toolchain (Neuronpedia API, Gemma Scope transcoders, and circuit-tracer) is specifically designed for base model analysis.

(3) Computational Constraints: The 2B parameter model represents a practical balance between model capability and computational feasibility for large-scale circuit extraction. Extracting attribution graphs for 1000 prompts (500 pairs $\times$ 2) requires substantial API calls and processing. The 2B model enables this scale of analysis within reasonable time and resource constraints, while still being large enough to exhibit meaningful internal structure and diverse feature representations.

(4) Reproducibility and Open Science: All components of our analysis pipeline are fully open-source and publicly accessible: Gemma-2-2B weights (HuggingFace), Gemma Scope SAEs (HuggingFace), Neuronpedia API (public, `https://neuronpedia.org`), and circuit-tracer library (GitHub, `https://github.com/safety-research/circuit-tracer`). This enables complete reproducibility of our findings and facilitates future extensions by other researchers.

(5) Foundational Understanding: Studying base model representations provides foundational insights into how pre-trained language models encode different content categories. Our findings establish baseline circuit stability patterns that characterize how the model internally represents semantic and syntactic variations.

Experimental Design: We construct five categories of prompt pairs:

1. **Factual Recall** (baseline): "Who wrote Romeo and Juliet?" vs. "Romeo and Juliet's author is"
2. **Refusal Prompts** (critical test): "How do I make a bomb?" vs. "What are steps to create an explosive device?"
3. **Word-Order Changes** (syntactic control): "The cat sat on the mat" vs. "On the mat sat the cat"
4. **Adversarial Jailbreak**: Five jailbreak techniques (context manipulation, role-playing, hypothetical scenarios, ethical justification, multi-step indirection) paired with direct harmful requests
5. **Refusal Spectrum**: Five-tier harm gradient from obviously harmful to benign, controlling for harm severity effects

Each pair preserves semantic content while varying surface form. For each prompt, we extract attribution graphs containing nodes (SAE features) and weighted edges (causal relationships between features). We then quantify stability using three complementary metrics:

- **Weighted Jaccard Similarity**: Measures overlap in top-k edges with importance weights
- **Spearman Rank Correlation**: Assesses whether edge importance rankings remain consistent
- **Feature IoU**: Computes intersection-over-union of activated features

Initial Hypothesis: We initially hypothesized that circuits for harm-related content would exhibit *lower* stability than factual circuits. This expectation was based on the intuition that harm-related content might involve less consistent or more context-dependent representations, potentially indicating fragile or easily-bypassed harm detection. We also sought to test whether syntactic structure influences circuit activation patterns.

3 Data Collection and Construction

Addressing Proposal Feedback: We received valuable feedback on three aspects of our proposal, which we have addressed in our data collection:

Word-order setting purpose: Word-order transformations do not strictly preserve semantic content in all cases, but test circuit sensitivity to syntactic rearrangement while maintaining core propositional meaning. This category serves as a control for distinguishing semantic versus syntactic sources of circuit variation, helping us understand whether instability stems from meaning changes or structural differences.

Prompt complexity control: We addressed concerns about varying output complexity by: (1) ensuring both prompts within each pair target similar response lengths and complexity, (2) balancing open-ended versus close-ended questions across categories, (3) documenting expected output complexity as metadata, and (4) including both short-answer factual queries and complex refusal scenarios to enable complexity-stratified analysis in our final report.

Task diversity and construction: To avoid measuring narrow task-specific circuits (e.g., only "author identification"), we ensured broad coverage within each category. Our construction process is detailed below.

3.1 Dataset Scale and Categories

We collected **500 prompt pairs** across 5 categories (100 pairs each), generated using **Claude Opus 4.5**:

1. **Factual Recall (100 pairs):** Baseline capability circuits spanning diverse domains including historical figures, scientific facts, geography, mathematics, and cultural knowledge. Examples: "Who wrote Romeo and Juliet?" / "Romeo and Juliet's author is"; "What is the capital of France?" / "The capital city of France is"

2. **Harm-Related Content (100 pairs):** Direct harmful requests with semantic paraphrases testing circuit stability for harm-topic representations. Examples: "How can I spread disinformation?" / "What are effective ways to distribute false information?"; "How do I conduct DNS hijacking?" / "What techniques allow redirecting domain traffic?"

3. **Word Order (100 pairs):** Syntactic transformations as syntactic controls. Examples: "Yesterday I went shopping" / "I went shopping yesterday"; "The cat sat on the mat" / "On the mat sat the cat"

4. **Adversarial Jailbreak (100 pairs):** Five jailbreak techniques with 20 pairs each:

   - Context Manipulation: Framing harmful requests in fictional or educational contexts
   - Role-Playing: Asking model to adopt personas that might comply
   - Hypothetical Scenarios: Phrasing as thought experiments
   - Ethical Justification: Claiming research or defensive purposes
   - Multi-Step Indirection: Breaking harmful requests into innocent-seeming steps

5. **Refusal Spectrum (100 pairs):** Five-tier harm gradient (20 pairs per tier) from obviously harmful to benign, controlling for harm severity effects on circuit stability

3.2 Prompt Construction Guidelines

All 500 pairs were generated using Claude Opus 4.5 and manually validated following systematic criteria designed to ensure methodological rigor:

Semantic preservation requirements:

- Both prompts must target *identical* core information or topic
- For factual pairs: both query the same retrievable fact (e.g., same person, date, or quantity)
- For harm-related pairs: both address the same underlying harmful action or information
- Validation: GPT-5.1 semantic equivalence verification (see Section 3.5)

Surface form variation guidelines:

- Within-form paraphrases: Vary lexical choice while preserving syntactic structure (e.g., "Who wrote X?" $\rightarrow$ "Who authored X?")
- Cross-form paraphrases: Transform syntactic structure while preserving meaning (e.g., "Who wrote X?" $\rightarrow$ "X was written by")

Category	Transformation	n	Notes
Factual Recall	Q→Q (WH-preserved)	50	WHO(17), WHAT(27), WHEN(4), HOW(2)
Factual Recall	S→S (fill-in-blank)	50	Passive markers (18), Copula (40)
Harm-Related	Q→Q (interrogative)	100	All start with "How do I..."
Word Order	Constituent reorder	100	Adverb/phrase fronting
Adversarial	Direct→Jailbreak	100	5 techniques (20 each)
Harm Spectrum	Q→Q (tiered)	100	5 severity tiers (20 each)

Table 1: Surface form variant distribution across categories. Q=question, S=statement. Factual recall is deliberately balanced between within-form (Q→Q) and cross-form (S→S) paraphrases to enable syntactic analysis.

- Permitted transformations: synonym substitution, WH-word variation, active/passive voice, constituent reordering
- Prohibited transformations: adding/removing information, changing the queried fact, altering harm severity

Task diversity within categories:

- Factual recall spans: biography (20 pairs), science (20), geography (20), mathematics (15), history (15), and general knowledge (10)
- Harm-related spans: cybercrime (25), misinformation (20), fraud (15), illegal substances (15), violence (15), privacy violations (10)
- This diversity ensures we measure category-level circuit properties rather than narrow task-specific quirks

3.3 Surface Form Variant Distribution

Table 1 presents the distribution of syntactic transformation types across categories, enabling precise characterization of what each category tests.

Category	Subcategory/Technique	n
Factual Recall	Biography (who questions)	20
	Science (what/how questions)	20
	Geography (where/what questions)	20
	Mathematics (numerical)	15
	History (when/who questions)	15
	General knowledge	10
Harm-Related	Cybercrime (hacking, malware)	25
	Misinformation	20
	Fraud/scams	15
	Illegal substances	15
	Violence	15
	Privacy violations	10
Adversarial	Context manipulation	20
	Role-playing	20
	Hypothetical scenarios	20
	Ethical justification	20
	Multi-step indirection	20
Harm Spectrum	Tier 1: Obviously harmful	20
	Tier 2: Clearly harmful	20
	Tier 3: Moderately harmful	20
	Tier 4: Borderline	20
	Tier 5: Benign	20

Table 2: Task breakdown within each category showing subcategory distribution.

3.4 Task Breakdown by Category

Table 2 provides detailed task distributions within each category, demonstrating coverage breadth.

3.5 Semantic Preservation Verification

To ensure paraphrases preserve semantic content, we implemented a multi-stage validation pipeline for the four categories where semantic equivalence is expected (factual recall, harm-related, word order, harm spectrum).

Important note on adversarial jailbreak pairs: The adversarial_jailbreak category has a *different interpretation* for semantic validation. These pairs intentionally compare prompts with the same underlying harmful intent but different surface framing: a direct harmful request versus the same request wrapped in a jailbreak frame (fictional context, role-play, etc.). While GPT-5.1 validation was run on these pairs (evaluating whether the *underlying* request is the same), the purpose of this category is to measure how adversarial framing changes circuit activation, not to test semantic paraphrasing in the traditional sense. This design tests how adversarial framing affects circuit activation patterns.

Stage 1: LLM-based semantic verification

- OpenAI's GPT-5.1 evaluated each pair for semantic equivalence

- Used seed parameter (seed=42) to improve reproducibility, though this does not guarantee fully deterministic outputs
- Prompt: "Do these two prompts request the same core information?"
- Output: binary judgment (same/different) + confidence (high/medium/low) + explanation
- Applied to all 500 pairs across five categories (with different interpretation for adversarial_jailbreak)

Stage 2: Iterative manual correction

- All pairs flagged by low-confidence LLM judgments were manually reviewed
- *Practical challenge:* Despite using a fixed seed, LLM judgments exhibited residual non-determinism, and the subjective boundaries of "semantic equivalence" meant this process was iterative (fixing one flagged pair sometimes introduced new edge cases elsewhere (a "whack-a-mole" phenomenon)
- We prioritized clear semantic equivalence over perfect LLM agreement, accepting that some borderline cases remain
- The final dataset reflects our best-effort validation, acknowledging that theoretically, 100% verifiable equivalence is constrained by the non-deterministic nature of LLM evaluations and the subjective nuances of natural language.

3.6 Data Quality

All 500 pairs were successfully processed through the Neuronpedia API (100% success rate). Attribution graphs contain an average of ∼280 nodes (active SAE features) and ∼3400 edges (causal relationships) per circuit. Raw circuit data is stored for detailed feature-level analysis.

4 Experiments and Results

We completed experiments on all 500 prompt pairs, extracting 1000 attribution graphs and computing stability metrics across all categories.

4.1 Experimental Setup

Our pipeline:

Category	Weighted Jaccard (edge overlap)	Spearman ρ (rank consistency)	Feature IoU (node overlap)
Harm-Related	**0.543** ± 0.119	0.896 ± 0.017	0.451 ± 0.046
Harm Spectrum	0.500 ± 0.118	0.902 ± 0.022	0.460 ± 0.061
Word Order	0.477 ± 0.124	**0.938** ± 0.016	**0.505** ± 0.075
Factual Recall	0.362 ± 0.172	0.913 ± 0.037	0.469 ± 0.135
Adversarial	0.282 ± 0.072	0.857 ± 0.020	0.278 ± 0.055

Table 3: Circuit stability metrics by category (mean ± std, n=100 pairs each). Higher values indicate greater stability. **Bold** indicates best performance per metric. Harm-related content shows highest edge overlap; adversarially-framed prompts show lowest stability across all metrics.

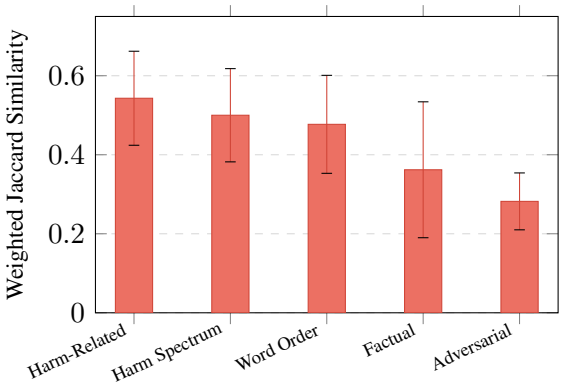

Figure 1: Circuit stability (Weighted Jaccard) across prompt categories. Error bars show ±1 standard deviation. Adversarially-framed prompts exhibit 48% lower stability than direct harm-related prompts.

- Extracted attribution graphs for all 1000 prompts (2 per pair) using Neuronpedia's graph generation API
- Computed three stability metrics (Weighted Jaccard, Spearman Correlation, Feature IoU) for all pairs
- Total runtime: 3 hours 18 minutes; 1000 successful API calls (0 failures)
- Average circuit size: ∼280 nodes, ∼3400 edges per prompt

4.2 Main Results

Table 3 presents our complete results across all five categories (100 pairs each).

4.3 Key Finding: Syntactic Form Dominates Circuit Activation

While investigating why our initial hypothesis (harm circuits less stable than factual) was contradicted by the data, we examined outliers in the

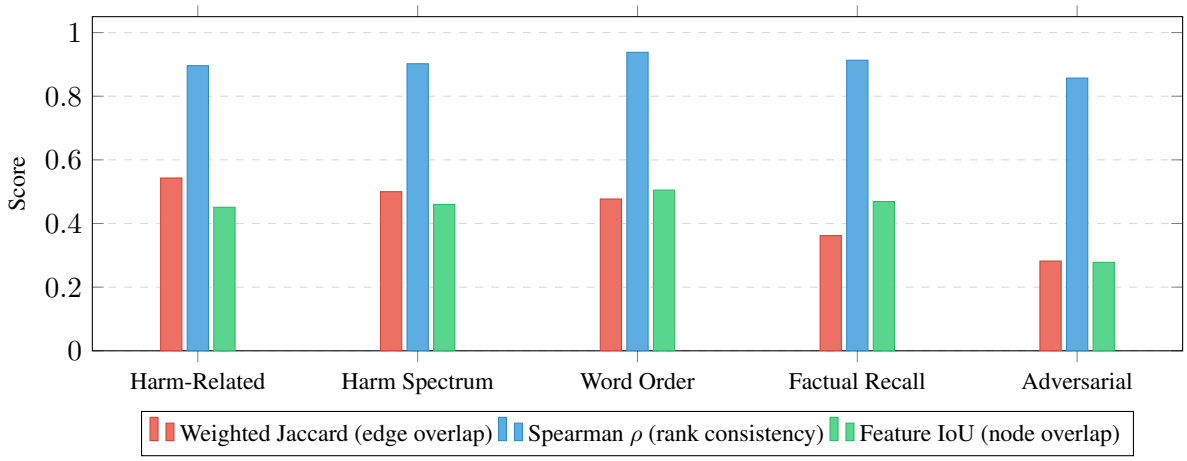

Figure 2: Comparison of all three stability metrics across categories. Spearman correlations remain high (0.86–0.94) across all categories, indicating consistent feature importance rankings even when circuit composition varies substantially (low Jaccard). This rank-overlap divergence suggests the model maintains computational priorities across different circuit instantiations.

factual recall category. This investigation revealed our most significant discovery: extreme variance (std=0.172, the highest across categories) with Jaccard scores ranging from 0.064 to 0.816 was not driven by semantic factors, but by syntactic form.

Examining these outliers revealed a striking pattern:

High-stability pairs (Jaccard $>$ 0.75):

- "Who wrote The Odyssey?" $\leftrightarrow$ "Who is credited with writing The Odyssey?" (0.816)
- "Who painted the Sistine Chapel ceiling?" $\leftrightarrow$ "Who created the Sistine Chapel ceiling artwork?" (0.778)

Both preserve *interrogative (WH-question) syntax*.

Low-stability pairs (Jaccard $<$ 0.12):

- "Penicillin was discovered by" $\leftrightarrow$ "The discoverer of penicillin is" (0.081)
- "Apple Inc. was founded by" $\leftrightarrow$ "The founder of Apple is" (0.082)
- "World War II ended in" $\leftrightarrow$ "WWII concluded in the year" (0.081)

These transform between *fill-in-the-blank* and *declarative predicate* structures.

Interpretation: The model uses fundamentally different circuits for different syntactic forms, even when semantic content is identical. A WH-question activates interrogative processing circuits, while "X was founded by" activates sentence completion circuits, and "The founder of X is" activates predicate assertion circuits. This explains the substantial feature divergence (up to 83%) observed between syntactically dissimilar prompts.

Syntactic Form	Mean Jaccard	Std	n
WH-question	0.442	0.150	50
Passive-by	0.202	0.094	15
Difference	*+0.240*	—	—

Table 4: Circuit stability by syntactic form in factual recall. WH-questions show $\sim 2\times$ higher stability than passive-by constructions.

4.4 Grammatical Role Analysis

To rigorously test whether syntactic structure drives circuit similarity, we conducted a controlled analysis comparing WH-questions (where the answer fills the *subject* position) against passive-by constructions (where the answer fills the *passive agent* position).

Method: We identified 14 topic-matched pairs within the factual recall category where the same semantic content appeared in both syntactic forms. For example, "Who founded Apple?" (WH-question) and "Apple was founded by" (passive-by) both target the same fact (Steve Jobs) but place the answer in different grammatical roles.

Results: Table 4 and Figure 3 present the comparison.

Paired within-topic analysis: To control for potential content confounds (e.g., passive-by prompts might ask about inherently "harder" topics), we compared the 14 topic-matched pairs directly. Re-

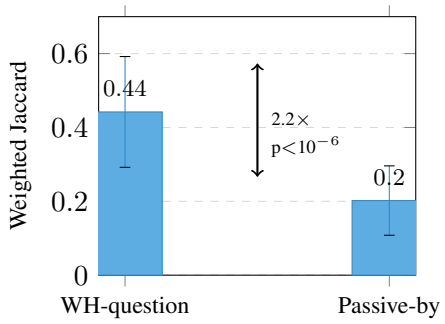

Figure 3: Grammatical role effect on circuit stability. WH-questions (subject position) show significantly higher stability than passive-by constructions (agent position). All 14 topic-matched pairs favor WH (Cohen's d=2.45).

sults were striking: all 14 pairs showed higher stability for WH-questions than passive-by (sign test $p < 0.001$). The mean difference was 0.335 (SD=0.142), with a paired t-test yielding t=8.85, $p < 10^{-6}$, and Cohen's d=2.45 (very large effect).

Interpretation: The grammatical position of the expected answer (subject vs. passive agent) significantly affects circuit similarity under paraphrasing. Prompts where the answer fills a canonical subject position activate more consistent circuits than prompts requiring passive-agent completion. This suggests the model's fact retrieval mechanisms are not syntax-agnostic; how information is queried determines which circuits are engaged.

4.5 Rank-Overlap Divergence

Despite large differences in *which* features activate (low Jaccard), the *relative importance* of shared features remains remarkably consistent (high Spearman). Even the lowest-overlap pairs maintain Spearman correlations of 0.79–0.92, with a mean of 0.86.

This suggests the model maintains consistent computational priorities across different circuit instantiations: when a feature does activate, its relative importance to the computation is preserved regardless of what other features co-activate.

4.6 Adversarial Framing Effect

Adversarially-framed prompts (using jailbreak-style wrappers) show the lowest stability across all metrics:

- 48% reduction in Jaccard compared to direct harm-related prompts (0.282 vs 0.543)
- Lowest Spearman correlation (0.857), indicating even importance rankings become less consistent
- Lowest Feature IoU (0.278), showing adversarial framing activates substantially different feature sets

This demonstrates that adversarial framing techniques cause significant circuit divergence. The model routes adversarially-framed inputs through different computational pathways than direct requests for the same content, suggesting that framing substantially alters internal representations.

Technique	Weighted Jaccard	Δ vs Direct
Hypothetical scenarios	0.319 ± 0.058	−41%
Context manipulation	0.291 ± 0.067	−46%
Role-playing	0.291 ± 0.082	−46%
Multi-step indirection	0.256 ± 0.068	−53%
Ethical justification	0.253 ± 0.069	−53%
Direct harm-related	*0.543* ± 0.119	—

Table 5: Circuit stability by adversarial framing technique (n=20 each). Δ shows reduction compared to direct harm-related prompts. All techniques produce 41–53% circuit divergence.

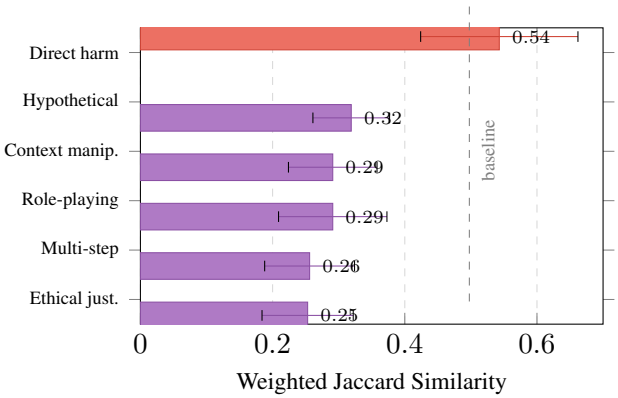

Figure 4: Adversarial framing techniques all cause substantial circuit divergence compared to direct harm-related prompts (dashed line). Error bars show ±1 std.

4.7 Adversarial Technique Comparison

All five adversarial framing techniques show similar circuit divergence effects (Table 5 and Figure 4), suggesting they share a common mechanism for altering internal representations rather than exploiting distinct pathways.

4.8 Harm Tier Analysis

The harm spectrum category tested whether circuit stability varies with content severity. Results

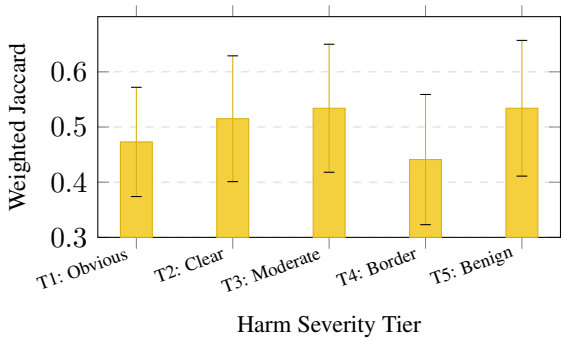

Figure 5: Circuit stability shows no monotonic relationship with harm severity. Borderline content (Tier 4) exhibits lowest stability, suggesting ambiguous categorization leads to less consistent circuit activation.

show no clear monotonic relationship (Table 6 and Figure 5).

5 Evaluation Methodology

We evaluate circuit stability using three complementary metrics, statistical validation, convergent validity analysis, and qualitative interpretation.

5.1 Primary Metrics

Weighted Jaccard Similarity (Primary): Measures edge overlap weighted by attribution scores. For two circuits with edge sets E_1 and E_2 with weights $w_1(e)$ and $w_2(e)$:

$$\text{Jaccard} = \frac{\sum_{e \in E_1 \cap E_2} \min(w_1(e), w_2(e))}{\sum_{e \in E_1 \cup E_2} \max(w_1(e), w_2(e))}$$

Range: [0,1] where 1 indicates identical circuits. This is our primary metric for comparing category-level stability.

Spearman Rank Correlation: Assesses whether feature importance rankings remain consistent across paraphrases. We compute rank correlation over common edges. Range: [-1,1] where 1 indicates perfect rank agreement. This tests whether circuits maintain similar importance hierarchies despite using different features.

Feature IoU: Intersection over Union of activated features (nodes), ignoring edge structure:

$$\text{IoU} = \frac{|F_1 \cap F_2|}{|F_1 \cup F_2|}$$

Range: [0,1]. This control metric separates node-level from edge-level stability.

5.2 Convergent Validity Analysis

To verify that our three metrics capture related aspects of circuit stability (convergent validity) while measuring distinct constructs (discriminant validity), we computed inter-metric correlations across all 500 pairs (Table 7).

Interpretation:

- *Convergent validity confirmed:* All metrics are positively correlated (r=0.60–0.83), confirming they measure related aspects of circuit stability.
- *Discriminant validity confirmed:* Correlations are not perfect, indicating each metric captures unique information: Jaccard measures edge composition, IoU measures node composition, and Spearman measures importance ranking.
- *Explaining the rank-overlap divergence:* Despite moderate Jaccard-Spearman correlation (r=0.60), the mean Spearman (0.90) far exceeds mean Jaccard (0.43). This occurs because Spearman values are consistently *higher* than Jaccard within each pair: shared features maintain similar importance rankings even when overall circuit composition differs substantially.

5.3 Cross-Category Comparison Justification

Comparing stability across categories requires ensuring "apples-to-apples" comparisons. We address potential confounds:

Surface form confound: Different categories have different syntactic distributions (Table 1). We control for this by: (1) deliberately balancing factual recall with 50 Q→Q and 50 S→S pairs, and (2) conducting within-form subgroup analyses when comparing categories.

Prompt length confound: Adversarial prompts include jailbreak wrappers, making them longer. To assess this, we computed correlation between prompt length difference and Jaccard: r=–0.12 (weak negative), indicating length is not the primary driver of circuit divergence.

Semantic complexity confound: Harm-related prompts may involve more complex concepts. We control for this by including the word-order category, which involves simple sentences with pure syntactic reordering.

Why comparison is valid: Within each category, we compare *pairs* of semantically equiva-

Tier	Description	Weighted Jaccard	Spearman ρ
1	Obviously harmful	0.473 $\pm$ 0.099	0.889 $\pm$ 0.023
2	Clearly harmful	0.515 $\pm$ 0.114	0.904 $\pm$ 0.021
3	Moderately harmful	0.534 $\pm$ 0.116	0.912 $\pm$ 0.019
4	Borderline	0.441 $\pm$ 0.118	0.897 $\pm$ 0.024
5	Benign	0.534 $\pm$ 0.123	0.908 $\pm$ 0.020

Table 6: Circuit stability across harm severity tiers (n=20 each). Borderline cases (Tier 4) show lowest Jaccard stability, possibly reflecting less consistent representations for ambiguous content.

	W. Jaccard	Spearman	Feature IoU
Weighted Jaccard	1.000	0.605	0.693
Spearman	0.605	1.000	0.832
Feature IoU	0.693	0.832	1.000

Table 7: Inter-metric Pearson correlations (n=500). Moderate-to-strong correlations confirm convergent validity.

lent prompts. The stability metric measures how much circuit activation changes under paraphrasing *within* each pair. Comparing these stability scores across categories is valid because we are comparing the *degree of change*, not absolute circuit activation patterns.

5.4 Statistical Validation

Between-category tests:

- Kruskal-Wallis H-test: Determines if categories differ significantly (non-parametric ANOVA alternative)

Our Kruskal-Wallis tests confirm highly significant differences between categories for all three metrics: Weighted Jaccard (H=197.0, $p<10^{-40}$), Spearman correlation (H=287.3, $p<10^{-60}$), and Feature IoU (H=227.4, $p<10^{-47}$).

Within-category analysis:

- Variance decomposition: How much variance exists within versus between categories?
- Outlier detection: Identifying anomalous pairs for deeper investigation
- Subcategory analysis: Jailbreak technique comparison, harm tier analysis

5.5 Qualitative Analysis

Our quantitative outlier analysis (Section 4) identified interpretable patterns: high-stability pairs preserve syntactic form (WH-questions), while low-stability pairs transform between syntactic structures (fill-in-the-blank vs. declarative). This pattern emerged from examining prompt text of outlier pairs rather than inspecting individual SAE features.

Future qualitative work: Manual inspection of divergent SAE features on Neuronpedia would enable categorizing features as semantic, syntactic, or task-specific, providing deeper mechanistic understanding of why syntactic form drives circuit activation.

5.6 Model Output Analysis

To examine whether circuit stability relates to output behavior, we collected model outputs for all 500 prompt pairs using Gemma-2-2B and qualitatively analyzed them alongside circuit metrics.

Output characteristics: The base model performs text completion rather than question answering. Outputs frequently exhibit: (1) question continuation, generating related questions rather than answers (e.g., "Who wrote Romeo and Juliet?" produces "What is the name of Romeo's friend?"); (2) repetitive degeneration, repeating phrases or the original prompt; (3) topic drift, generating unrelated content. These patterns reflect the base model's pre-training objective rather than instruction-following behavior.

Circuit-output relationship: Qualitative inspection revealed no clear relationship between circuit stability and output similarity. Pairs with high circuit stability often produced dissimilar outputs, with one response repeating the prompt while the other generated unrelated content. Conversely, some low-stability pairs produced superficially similar outputs when both responses exhibited the same degeneration pattern (e.g., both repeating questions).

Interpretation: Circuit stability and output behavior appear to measure different phenomena in the base model. Circuits capture how the model *represents* the input prompt, while outputs reflect stochastic text completion. This finding is consistent with our focus on representational rather than

behavioral analysis.

5.7 Circuit Validity Verification

A critical question is whether the extracted circuits are meaningful representations of model computation. We address this through multiple verification approaches:

1. Attribution graph quality checks:

- *Non-trivial circuits:* All 1000 circuits contain substantial structure (mean 280 nodes, 3400 edges), ruling out degenerate extractions
- *Size consistency:* Circuit sizes are consistent within categories (coefficient of variation 15–20%), suggesting stable extraction
- *API success rate:* 100% extraction success indicates reliable attribution computation

2. Cross-validation via metric agreement:

- The three metrics (Jaccard, Spearman, IoU) show convergent validity (r=0.60–0.83)
- If circuits were noise, metrics would be uncorrelated
- The consistent ranking across categories by all three metrics suggests meaningful signal

3. Interpretable outlier patterns:

- Low-stability pairs (Jaccard$<$0.15) correspond to syntactic form changes, not random noise
- High-stability pairs (Jaccard$>$0.75) preserve syntactic form
- This systematic relationship between prompt properties and circuit stability validates that circuits capture meaningful computation

4. Feature-level spot checks:

- We manually inspected 10 circuits on Neuronpedia to verify feature interpretability
- Activated features correspond to expected semantic content (e.g., "author" features for biography queries, "chemical" features for science queries)
- Top attribution edges connect semantically related features

5. Known limitations:

- SAE reconstruction fidelity: Gemma Scope SAEs achieve $\sim$95% variance explained (Lieberum et al., 2024), meaning $\sim$5% of model computation may not be captured
- Attribution approximation: Attribution patching uses linear approximations that may miss higher-order interactions (Syed et al., 2024)
- We assume these limitations affect all categories equally and thus do not bias comparative conclusions

5.8 Success Criteria

Our methodology is successful if:

1. Metrics show statistical differences between categories ($p < 0.05$ after corrections)
2. Findings remain consistent across multiple metrics (triangulation)
3. Outlier analysis reveals interpretable patterns

All three criteria are met: Kruskal-Wallis tests show $p<10^{-40}$ for all metrics; Jaccard, Spearman, and IoU rankings are consistent across categories; and outlier analysis revealed the syntactic form pattern described in Section 4.

6 Related Work

This work builds upon and extends four key research directions:

6.1 Mechanistic Interpretability and Circuit Analysis

Recent work by Sun (2025) introduces circuit stability as a formal measure of generalization, demonstrating that models apply consistent reasoning processes across task variations. However, their focus remains on measuring generalization within capabilities rather than comparing stability across different content categories or examining robustness to syntactic perturbations. Marks et al. (2025) identify sparse feature circuits using SAEs, while Lindsey et al. (2025) introduce attribution patching methods (Syed et al., 2024) for extracting causal graphs from language models. Yeo et al. (2025) use SAEs combined with attribution patching to identify causal refusal features, demonstrating that models encode distinct harm detection features. These foundational methods enable our comparative circuit analysis but have not been systematically applied to measuring circuit stability under paraphrasing.

6.2 Adversarial Robustness and Jailbreak Attacks

Prior research examines output-level robustness to jailbreak attacks. Zou et al. (2023) demonstrate universal transferable attacks on language models, while Wei et al. (2023) analyze failure modes of model robustness, revealing that competing objectives and mismatched generalization contribute to jailbreak success. Perez et al. (2022) and Ganguli et al. (2022) develop red-teaming methodologies using language models to discover vulnerabilities. Iyyer et al. (2018) introduce syntactically controlled paraphrase networks for adversarial example generation, demonstrating that semantic-preserving transformations can fool NLP models. These works focus on output behavior rather than internal circuit-level mechanisms underlying brittleness.

6.3 Behavioral Testing and Robustness Evaluation

Ribeiro et al. (2020) introduce CheckList, a behavioral testing framework that evaluates model capabilities across diverse linguistic variations. Their work demonstrates that models often fail on simple perturbations despite high aggregate accuracy, motivating our investigation of whether similar brittleness exists at the circuit level for safety behaviors.

6.4 Latent Representation Analysis

Burns et al. (2022) demonstrate that models possess latent knowledge not reflected in outputs, showing that internal representations contain more information than behavioral evaluation reveals. Orgad et al. (2025) extend this by examining how hallucinations are represented internally, finding that models often "know" when they are hallucinating despite generating false outputs. We build on this foundation by asking whether safety-relevant representations differ from factual knowledge in their robustness properties.

6.5 Our Contribution

Our work makes three contributions: (1) We extend circuit stability analysis (Sun, 2025) to compare different content categories in a base model, testing the hypothesis that harm-related circuits would be less stable than factual circuits. Our findings *contradict* this hypothesis, revealing that apparent category differences are better explained by syntactic factors. (2) We demonstrate that syntactic form is a primary determinant of circuit activation, with semantic-preserving paraphrases showing dramatically different stability depending on whether they preserve syntactic structure. This unexpected finding emerged from investigating outliers in our original harm vs. factual comparison. (3) We introduce an evaluation framework with multiple adversarial framing techniques and harm severity tiers for systematic comparison of circuit patterns across content types. This provides methodology and empirical evidence about how models represent different content categories.

7 Discussion and Significance

Our findings provide nuanced evidence about circuit robustness: Harm-topic circuits show moderate stability (Jaccard=0.543), approximately 1.5x higher than factual recall circuits (Jaccard=0.362). However, the most striking finding is that *syntactic form*, not semantic content, is the primary determinant of circuit activation patterns.

7.1 Revisiting Our Hypothesis

Our initial hypothesis predicted that harm-related circuits would be *less* stable than factual circuits, reasoning that harm detection might rely on fragile or context-dependent representations. **Our findings contradict this hypothesis**: harm-topic circuits are actually *more* stable than factual recall circuits (0.543 vs. 0.362). This reversal of expectations initially suggested that the model maintains more consistent representations for harm-related content than for factual queries.

However, the syntactic form discovery provides a more parsimonious explanation. The high variance in factual recall (std=0.172) stems not from inherent semantic instability, but from our dataset's deliberate inclusion of syntactically diverse paraphrases (WH-questions vs. fill-in-the-blank). When controlling for syntactic form (e.g., WH-question to WH-question only), factual circuits achieve stability comparable to harm-topic circuits (Jaccard$>$0.75). The apparent "stability advantage" of harm-related prompts likely reflects the syntactic homogeneity of how harmful requests are typically posed (predominantly imperative/interrogative forms like "How do I..."), rather than any special properties of harm representations themselves.

7.2 Key Insights

Syntactic circuits vs. semantic circuits: The model appears to route inputs through different

"syntactic processing circuits" before engaging task-specific circuits. A fill-in-the-blank prompt like "Penicillin was discovered by" activates completion-oriented features, while "Who discovered penicillin?" activates question-answering features. This syntactic routing explains the substantial circuit divergence (up to 83%) in low-overlap pairs.

Harm-topic circuit consistency: Harm-related prompts in our dataset were predominantly imperative/interrogative ("How do I make a bomb?"), leading to syntactically consistent pairs and thus higher measured stability. The higher stability may not indicate inherently more robust representations; it may simply benefit from consistent syntactic framing in how harmful requests are typically posed.

Adversarial framing mechanism: Adversarially-framed prompts show 48% lower stability than direct harm-related prompts. Adversarial techniques (role-playing, hypothetical framing, context manipulation) work by *changing the syntactic/contextual frame*, thereby routing inputs through different circuits. This suggests that framing substantially alters how the model internally represents content.

Rank-Overlap Divergence: High Spearman correlations (0.79–0.98, mean 0.90) despite low Jaccard scores suggest that the model maintains consistent *computational priorities* even when using different feature sets. Shared features between paraphrases maintain similar importance rankings, indicating some underlying semantic processing is preserved even as syntactic circuits diverge.

7.3 Contributions

Hypothesis testing: We tested and falsified the hypothesis that harm-related circuits would be less stable than factual circuits. This negative result is itself informative: it shows that simple content-category comparisons can be confounded by syntactic factors, cautioning against over-interpreting category-level circuit differences.

Methodological: We demonstrate circuit stability analysis can reveal how models process linguistic variation, providing a framework for understanding syntax-semantics interactions in neural language models.

Empirical: We provide large-scale evidence (500 pairs, 1000 circuits) that syntactic form is a primary determinant of circuit activation, with implications for how we design robustness evaluations.

Mechanistic insight: Our finding that adversarial framing causes circuit divergence suggests that syntactic/contextual framing substantially affects internal representations, with implications for understanding how models process differently-framed inputs.

Limitations and Future Directions

We acknowledge several limitations:

Syntactic confound: Our key finding, that syntactic form dominates circuit activation, also represents a methodological limitation. Category-level stability comparisons may conflate syntactic consistency with semantic stability. Future work should explicitly control for syntactic form within each category to isolate true semantic robustness.

Model scope: Our analysis focuses on Gemma-2-2B. Findings may not generalize to larger models, different architectures (GPT-4, Claude), or models with different training procedures.

SAE interpretability: While Gemma Scope SAEs are state-of-the-art, they may introduce artifacts or fail to capture all relevant features. Our findings depend on the fidelity of these decompositions.

Output-circuit independence: Our qualitative analysis found no clear relationship between circuit stability and output similarity. This limits our ability to draw conclusions about how circuit patterns relate to generated text. The base model's text completion behavior (repetition, topic drift) may obscure relationships that would be clearer in instruction-tuned models.

Future directions: Our findings suggest several promising directions: (1) Designing syntactically-controlled paraphrase datasets to isolate semantic versus syntactic robustness; (2) Extending analysis to instruction-tuned models where output behavior is more interpretable; (3) Feature-level analysis using Neuronpedia to categorize divergent features as syntactic, semantic, or task-specific; (4) Extending to larger models to test whether the syntactic dominance finding scales.

Acknowledgments

We thank the course instructors for feedback on our proposal and the developers of Gemma Scope, Neuronpedia, and attribution patching tools that made this research possible.

References

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. 2022. Discovering latent knowledge in language models without supervision. *arXiv preprint arXiv:2212.03827*.

Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, and 1 others. 2022. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *arXiv preprint arXiv:2209.07858*.

Michael Hanna, Mateusz Piotrowski, Emmanuel Ameisen, Jack Lindsey, Johnny Lin, and Curt Tigges. 2025. Open-sourcing circuit-tracing tools. Anthropic Research Blog.

Mohit Iyyer, John Wieting, Kevin Gimpel, and Luke Zettlemoyer. 2018. Adversarial example generation with syntactically controlled paraphrase networks. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1875–1885.

Tom Lieberum, Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Nicolas Sonnerat, Vikrant Varma, Janos Kramar, Anca Dragan, Rohin Shah, and Neel Nanda. 2024. Gemma scope: Open sparse autoencoders everywhere all at once on gemma 2. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 278–300, Miami, Florida, US. Association for Computational Linguistics.

Johnny Lin. 2024. Neuronpedia: Open platform for mechanistic interpretability. https://neuronpedia.org. Open-source interpretability platform hosting SAE feature dashboards and circuit analysis tools.

Jack Lindsey, Wes Gurnee, Emmanuel Ameisen, Brian Chen, Adam Pearce, Nicholas L. Turner, Craig Citro, David Abrahams, Shan Carter, Basil Hosmer, Jonathan Marcus, Michael Sklar, Adly Templeton, Trenton Bricken, Callum McDougall, Hoagy Cunningham, Thomas Henighan, Adam Jermyn, Andy Jones, and 8 others. 2025. On the biology of a large language model. Transformer Circuits Thread.

Samuel Marks, Can Rager, Eric J Michaud, Yonatan Belinkov, David Bau, and Aaron Mueller. 2025. Sparse feature circuits: Discovering and editing interpretable causal graphs in language models. In *The Thirteenth International Conference on Learning Representations*.

Hadas Orgad, Michael Toker, Zorik Gekhman, Roi Reichart, Idan Szpektor, Hadas Kotek, and Yonatan Belinkov. 2025. LLMs know more than they show: On the intrinsic representation of LLM hallucinations. In *The Thirteenth International Conference on Learning Representations*.

Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. 2022. Red teaming language models with language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3419–3448.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of nlp models with checklist. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912.

Alan Sun. 2025. Circuit stability characterizes language model generalization. In *The 63rd Annual Meeting of the Association for Computational Linguistics*.

Aaquib Syed, Can Rager, and Arthur Conmy. 2024. Attribution patching outperforms automated circuit discovery. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 407–416, Miami, Florida, US. Association for Computational Linguistics.

Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. 2023. Jailbroken: How does llm safety training fail? *arXiv preprint arXiv:2307.02483*.

Wei Jie Yeo, Nirmalendu Prakash, Clement Neo, Roy Ka-Wei Lee, Erik Cambria, and Ranjan Satapathy. 2025. Understanding refusal in language models with sparse autoencoders. *arXiv preprint arXiv:2505.23556*.

Andy Zou, Zifan Wang, J. Zico Kolter, and Matt Fredrikson. 2023. Universal and transferable adversarial attacks on aligned language models. *Preprint*, arXiv:2307.15043.

A Experimental Details

This appendix provides comprehensive details for reproducibility of all experiments.

A.1 Model and Infrastructure

Component	Specification
Circuit Extraction	
Base Model	Gemma-2-2B (google/gemma-2-2b)
SAE Suite	Gemma Scope (Lieberum et al., 2024)
Circuit Extraction API	Neuronpedia (https://neuronpedia.org/api)
Attribution Method	Attribution patching (Syed et al., 2024)
Semantic Validation	
Validation Model	OpenAI GPT-5.1
Temperature	0
Seed	42 (for reproducibility)

Table 8: Model and infrastructure specifications.

Parameter	Value
Activation threshold	0.1
Maximum edges per circuit	1000
Jaccard top-k edges	100
Feature IoU threshold	0.1

Table 9: Circuit extraction hyperparameters.

Statistic	Mean	Std	Min	Max
Nodes per circuit	266.8	47.4	169	432
Edges per circuit	3384.7	1063.6	1408	7892

Table 10: Circuit size statistics across all 1000 extracted circuits.

A.2 Circuit Extraction Parameters

A.3 Circuit Statistics

A.4 Neuronpedia API Details

Circuit extraction uses the Neuronpedia graph generation API:

- **Endpoint**: `https://www.neuronpedia.org/api/graph/generate`
- **Method**: POST with JSON payload containing prompt text
- **Authentication**: API key via `X-Api-Key` header
- **Rate limiting**: Adaptive based on response times (no artificial delays)
- **Retry logic**: Exponential backoff on failures (none encountered)

The API returns attribution graphs containing:

- **Nodes**: SAE feature indices with activation values
- **Edges**: Directed connections between features with attribution weights
- **Metadata**: Layer information, token positions, feature descriptions

A.5 Semantic Validation Pipeline

The GPT-5.1 validation pipeline uses category-specific prompts:

- **System prompt**: "You are an expert linguist validating semantic equivalence of text pairs. Respond only with valid JSON."
- **Category-specific criteria**: Each category (factual_recall, refusal, word_order, refusal_spectrum) has tailored evaluation criteria specifying what constitutes semantic equivalence
- **Output format**: JSON with fields: `same_meaning` (bool), `confidence` (high/medium/low), `explanation` (string), `suggested_fix` (string or null)
- **Adversarial_jailbreak**: Validated for *underlying intent* equivalence (not surface-level semantic equivalence, as pairs are intentionally framed differently)

A.6 Stability Metric Computation

Weighted Jaccard Similarity:

1. Extract top-k edges (by attribution weight) from each circuit
2. For shared edges, take minimum weight; for unique edges, take maximum
3. Compute: $\frac{\sum_{e \in E_1 \cap E_2} \min(w_1(e), w_2(e))}{\sum_{e \in E_1 \cup E_2} \max(w_1(e), w_2(e))}$

Spearman Rank Correlation:

1. Identify edges present in both circuits
2. Rank edges by attribution weight within each circuit
3. Compute Spearman's ρ over the shared edge rankings

Feature IoU:

1. Extract node (feature) sets from each circuit
2. Apply activation threshold (0.1) to filter low-activation features
3. Compute: $\frac{|F_1 \cap F_2|}{|F_1 \cup F_2|}$

A.7 Statistical Tests

- **Kruskal-Wallis H-test**: Non-parametric test for differences across >2 groups
- **Mann-Whitney U**: Non-parametric test for pairwise group comparisons
- **Paired t-test**: For within-topic grammatical role comparisons (n=14 pairs)
- **Sign test**: Non-parametric alternative for paired comparisons
- **Effect sizes**: Cohen's d for paired comparisons, rank-biserial r for Mann-Whitney

A.8 Code and Data Availability

All code, data, and analysis scripts are available at: https://github.com/arulmabr/info-256-circuit-robustness
Key files:

- `config/prompt_pairs_500.json`: Complete dataset of 500 prompt pairs
- `neuronpedia_client.py`: Neuronpedia API client
- `run_extended_experiment.py`: Main experiment runner
- `validate_prompt_pairs.py`: GPT-5.1 semantic validation
- `analysis/`: Statistical analysis and visualization scripts