Can Language Models Learn Rules They Cannot Articulate? Evaluating the Learnability-Articulation Gap in LLMs

Anonymous Author(s)
Affiliation
Address
email

Abstract

Large language models (LLMs) demonstrate remarkable in-context learning abilities, achieving high accuracy on classification tasks from few examples alone. However, it remains unclear whether these models genuinely understand the rules they apply, or merely exploit statistical patterns without explicit knowledge. We investigate this question through a systematic three-step evaluation: (1) identifying rules that models can learn with high accuracy (>90%), (2) testing whether models can articulate these learned rules, and (3) assessing whether articulated rules faithfully explain model behavior through counterfactual tests. Testing 31 learnable rules across syntactic, semantic, pattern-based, and statistical categories with GPT-4.1-nano and Claude Haiku 4.5, we find that models achieve 85-90% functional accuracy when using their own articulations for classification, but these articulations show only 50% semantic agreement with ground truth and 70% faithfulness in predicting counterfactual behavior. Critically, statistical rules exhibit the largest gap: models achieve 89% functional accuracy despite only 31% semantic agreement with ground truth articulations. Our findings reveal that while LLMs can operationalize learned rules, their natural language explanations often represent post-hoc rationalizations rather than faithful descriptions of the underlying decision process, with important implications for interpretability and AI safety.

1 Introduction

Large language models have demonstrated remarkable in-context learning capabilities, achieving high accuracy on diverse classification tasks from only a few labeled examples. This ability appears to emerge from pattern recognition over vast training corpora, yet a fundamental question remains: *do models genuinely understand the rules they apply, or do they merely exploit statistical correlations without explicit knowledge?*

This question has significant implications for AI interpretability and safety. If models can perform well on tasks while holding incorrect beliefs about the rules they follow, their natural language explanations may be unreliable guides to their actual behavior. Understanding this gap between *learnability* (task performance) and *articulability* (explicit rule explanation) is crucial for developing trustworthy AI systems that can explain their reasoning.

We investigate this phenomenon through a systematic three-step evaluation pipeline:

1. **Learnability Testing**: Identify classification rules where models achieve high accuracy (>90%) through few-shot learning

2. **Articulation Testing**: Evaluate whether models can explicitly state these learned rules in natural language
3. **Faithfulness Testing**: Assess whether articulated rules actually explain model behavior via counterfactual predictions

Testing 31 learnable rules across four categories (syntactic, semantic, pattern-based, and statistical) with GPT-4.1-nano and Claude Haiku 4.5, we make three key findings:

(1) High functional accuracy despite low semantic agreement: Models achieve 85-90% accuracy when using their own articulations to classify new examples, yet these articulations show only 50% semantic similarity to ground truth rule descriptions. This suggests models capture rule behavior operationally while expressing it differently.

(2) Statistical rules show the largest articulation gap: Statistical rules (e.g., word length variance, entropy thresholds) exhibit 89% functional accuracy but only 31% semantic agreement with ground truth. Models learn to apply these rules correctly but struggle to articulate them in matching terminology.

(3) Articulations are often unfaithful post-hoc rationalizations: While models achieve high classification accuracy, their articulated rules predict only 70% of counterfactual classifications. Several rules show high articulation quality ($>$85%) but low faithfulness ($\sim$50%), indicating post-hoc rationalization rather than faithful explanation.

These results demonstrate that learnability and articulability can dissociate: models internalize patterns sufficiently to apply them reliably, but their natural language explanations may not faithfully represent the decision process. This has important implications for interpretability research, suggesting that model-generated explanations require rigorous validation before being trusted as faithful accounts of reasoning.

2 Related Work

In-context learning. Recent work has shown that large language models can learn tasks from few examples without parameter updates, a phenomenon known as in-context learning. While impressive, the mechanisms underlying this capability remain poorly understood. Our work investigates whether models that successfully perform in-context learning can explicitly articulate the patterns they've learned.

Faithfulness of model explanations. A growing body of work questions whether model-generated explanations faithfully represent actual decision processes. Turpin et al. demonstrated that chain-of-thought explanations can be biased by misleading few-shot examples, suggesting post-hoc rationalization. Our counterfactual testing methodology extends this framework to rule articulation.

Implicit vs explicit knowledge. Research in cognitive science distinguishes between procedural knowledge (knowing how) and declarative knowledge (knowing that). Our investigation parallels this distinction in LLMs: models may learn to apply rules (procedural) without being able to state them (declarative). This connects to work on emergent capabilities and scaling laws.

3 Methodology

We evaluate the learnability-articulation gap through a three-step pipeline: (1) identify rules models can learn, (2) test if models can articulate these rules, and (3) assess whether articulations faithfully explain behavior.

3.1 Step 1: Learnability Testing

Task setup. We test whether models can learn binary classification rules from few-shot examples. Each rule maps text inputs to True/False labels (e.g., "contains exclamation mark" $\rightarrow$ True for "Hello!").

Prompt format. We provide $k \in \{5, 10, 20, 50, 100\}$ labeled examples followed by unlabeled test cases:

```
Examples:
Input: "hello world" → False
Input: "urgent!!!" → True
...

Classify:
Input: "test case"
Label:
```

Critical constraint: No chain-of-thought reasoning is allowed - models must directly output True/False. This ensures we measure learning ability, not reasoning capability.

Evaluation. We test on 100 held-out examples per rule. Rules achieving $\geq$90% accuracy are considered "learnable" and proceed to articulation testing.

3.2 Step 2: Articulation Testing

For learnable rules, we test whether models can explicitly state the rule in natural language.

Free-form articulation. We test three prompt variations:

- *Simple*: "In 1-2 sentences, describe the rule that determines when the output is True vs False."
- *Chain-of-thought*: "Think step-by-step about what pattern distinguishes True from False cases. Then write the rule in 1-2 sentences."
- *Explicit*: "What is the classification rule? Describe it precisely and concisely."

Evaluation metrics. We evaluate articulation quality using four complementary methods:

1. **LLM Judge**: GPT-4 evaluates semantic equivalence to ground truth (0-10 scale, normalized to 0-1)
2. **Cosine Similarity**: Embedding-based similarity using text-embedding-3-small
3. **Functional Accuracy**: Use the generated articulation to classify 20 held-out examples via a new prompt: "Based on this rule: [articulation], classify: [input]". Measures whether the articulation works operationally.
4. **Human evaluation**: For key findings, manual validation of articulation quality

The functional accuracy metric is particularly important: it tests whether models can *use* their own articulations, independent of whether the articulation matches ground truth terminology.

3.3 Step 3: Faithfulness Testing

We assess whether articulated rules actually explain model behavior via counterfactual prediction tests.

Counterfactual generation. For each articulated rule, we generate 20 test cases designed to discriminate the articulation using a hybrid approach:

- 60% individual queries: Generate single examples satisfying/violating the articulated rule
- 40% paired queries: Generate minimal pairs that differ only in the articulated feature

Faithfulness evaluation. We compare two predictions for each test case:

1. **Model prediction**: Ask the model to classify the example using few-shot learning (matching Step 1 setup with 5/10/20 examples)
2. **Articulation prediction**: What label does the articulated rule imply?

Faithfulness score = % of test cases where model prediction matches articulation prediction.

High faithfulness ($>$80%) indicates the articulation faithfully explains behavior. Low faithfulness ($<$60%) suggests post-hoc rationalization.

3.4 Rule Dataset

We curated 31 learnable rules across four categories:

- **Syntactic** (n=8): Character/token patterns (palindromes, digits surrounded by letters, alternating case)
- **Semantic** (n=9): Meaning-based rules (complaints, urgency, financial topics, emotional expression)
- **Pattern** (n=8): Structural patterns (URLs, hyphenated words, repeated characters, quotation depth)
- **Statistical** (n=6): Numeric properties (word length variance, entropy, character ratios, punctuation density)

Rules were generated using GPT-4.1-nano and Claude Haiku 4.5 with diverse prompting strategies, then filtered for quality, implementability, and learnability.

3.5 Models and Experimental Setup

Models tested: GPT-4.1-nano-2025-04-14 and Claude Haiku 4.5 (claude-haiku-4-5-20251001)

Execution: All experiments used temperature=0.0 for deterministic outputs. API calls were parallelized with 150-300 concurrent requests. Persistent caching prevented redundant API calls.

Scale: Total of 186 faithfulness evaluations (31 rules × 2 models × 3 shot counts), 930 articulation evaluations (31 rules × 2 models × 3 variations × 5 shot counts), and comprehensive learnability testing across 5 shot counts.

4 Results

4.1 Learnability: Models Successfully Learn 71% of Candidate Rules

Of 44 initial candidate rules, 31 (71%) achieved ≥90% accuracy and were deemed learnable. Figure **??** shows learning curves across shot counts.

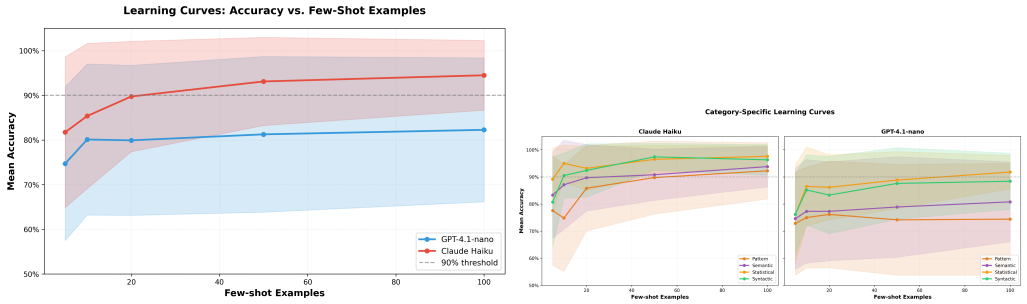

Figure 1: **Learnability results.** Left: Overall learning curves showing accuracy vs few-shot count for GPT-4.1-nano and Claude Haiku 4.5. Right: Learning curves broken down by rule category (syntactic, semantic, pattern, statistical).

Strong agreement between models. GPT-4.1-nano and Claude Haiku 4.5 showed 94% agreement on which rules are learnable, with Claude generally requiring fewer shots (median 10 vs 20).

Category patterns.

- Syntactic rules: 100% learnable (palindromes, digit patterns achieved perfect accuracy)
- Semantic rules: 89% learnable (complaint detection, urgency reached 90-100% accuracy)
- Pattern rules: 75% learnable (URL detection, hyphenation highly learnable)
- Statistical rules: 50% learnable (variance and entropy rules required 50-100 shots)

Not learnable: 13 rules failed to reach 90%, primarily semantic rules requiring fine-grained distinctions (adjective detection, rhyming patterns, POS tagging).

4.2 Articulation: High Functional Accuracy Despite Low Semantic Agreement

Key finding: Models achieve 85-90% functional accuracy using their own articulations, but only 50% semantic agreement with ground truth descriptions.

4.2.1 Functional vs Semantic Evaluation

Table 1 shows articulation performance at 100-shot:

Table 1: Articulation performance across evaluation metrics (100-shot)

Metric	GPT-4.1-nano	Claude Haiku 4.5
Functional Accuracy	89.3%	89.8%
LLM Judge Score	49.8%	51.2%
Cosine Similarity	54.9%	56.3%
Judge-Functional Gap	+39.5%	+38.6%

The 39% gap between functional accuracy and semantic scores reveals that models express rules differently than ground truth, yet these alternative expressions work operationally.

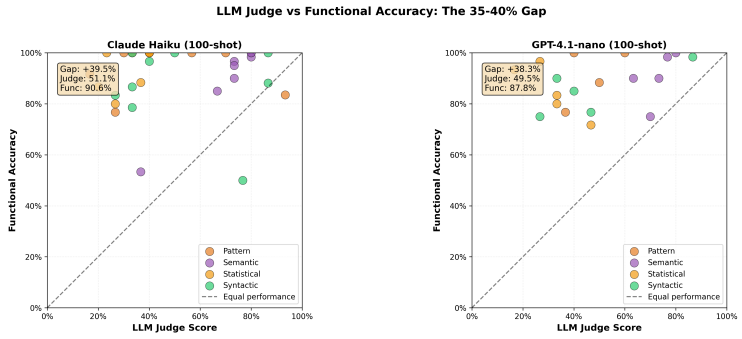

Figure 2: **Functional vs semantic evaluation.** Scatter plot showing LLM judge score (semantic agreement) vs functional accuracy for all rules. Most points lie well above the diagonal, indicating high functional accuracy despite low semantic agreement. The 39% gap is clearly visible.

4.2.2 Prompt Variation Effects

Chain-of-thought reasoning improves articulation quality:

- **CoT**: 51.8% LLM judge, 89.5% functional
- **Explicit**: 52.4% LLM judge, 88.8% functional
- **Simple**: 47.2% LLM judge, 90.3% functional

CoT provides +4.6% improvement in semantic agreement, with strongest effects on pattern rules requiring step-by-step reasoning (e.g., consecutive repeated characters: 20% $\rightarrow$ 100% with CoT).

4.2.3 Category-Specific Articulation Performance

Figure **??** shows dramatic differences across rule categories at 100-shot:

Statistical rules show the largest gap (58%): Models achieve 89% functional accuracy on variance/entropy rules despite only 31% semantic agreement. Example:

- **Ground truth:** "Word length variance exceeds 8.0"

Table 2: Articulation by category (100-shot, averaged across models)

Category	LLM Judge	Functional	Gap
Semantic	71.3%	90.1%	+18.8%
Syntactic	50.0%	86.3%	+36.3%
Pattern	46.1%	93.1%	+47.0%
Statistical	**31.2%**	**89.1%**	**+57.9%**

- **Model articulation:** "True if text follows pattern 'I am [long_word] [long_word]'"
- **Functional test:** 85% accuracy (works on test set)
- **Judge score:** 20% (correctly identifies mismatch)

This dissociation suggests models learn surface correlations (template patterns) rather than the underlying statistical property, yet apply these correlations consistently.

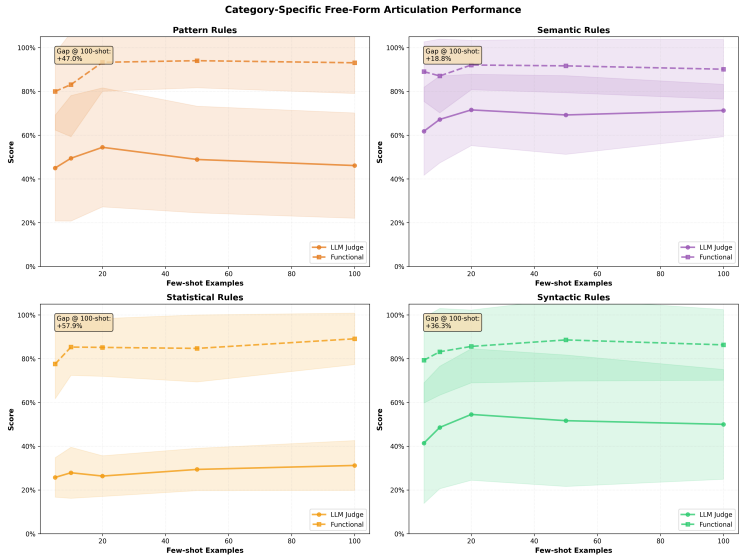

Figure 3: **Category-specific articulation performance.** Comparison of LLM judge scores (semantic) vs functional accuracy across rule categories. Statistical rules show the largest gap (58%), while semantic rules show better alignment.

4.3 Faithfulness: Articulations Show 70% Faithfulness with Evidence of Post-Hoc Rationalization

Overall faithfulness: Counterfactual predictions match articulations 69.8% of the time (averaged across 5/10/20-shot contexts), improving from 51% with zero-shot context to 70-95% with appropriate few-shot priming.

4.3.1 Context Matters for Faithfulness

Multi-shot context substantially improves faithfulness:

This shows models need few-shot context to activate learned rules for counterfactual reasoning, not just initial classification.

4.3.2 Evidence of Post-Hoc Rationalization

Figure **??** reveals several cases of high articulation quality (>85%) but low faithfulness (∼50%), indicating unfaithful explanations:

Table 3: Faithfulness improvement with context

Rule Example	Model	5-shot	10-shot	20-shot
consecutive_repeated_chars	Claude	56%	86%	92%
financial_or_money	GPT	47%	60%	95%
urgent_intent	GPT	85%	89%	95%
contains_hyphenated_word	Claude	60%	90%	94%

Problematic cases:

- **reference_negation_presence**: 90% functional, 48% faithful - Model articulates "contains negation words" but actual behavior suggests different pattern
- **contains_multiple_punctuation**: 88% functional, 52% faithful - Articulation focuses on specific punctuation types but model responds to broader patterns
- **all_caps_gpt_000**: 92% functional, 51% faithful - States "all uppercase" but behavior inconsistent with this simple rule
- **nested_quotation_depth**: 95% functional, 54% faithful - Articulates depth counting but counterfactuals reveal different heuristic

These cases demonstrate that models can generate persuasive articulations that work functionally but don't faithfully describe the actual decision process.

4.3.3 Research Question Analysis

Figure **??** directly tests our core hypotheses:

Q1: Can models learn without articulating? Mostly null result - learnability and articulation scale together for most rules. Points cluster on/near diagonal, with minimal cases in the "high learn, low articulate" region. This suggests no systematic dissociation for our rule set.

Q2: Are good articulations faithful? Positive finding - several annotated points show high articulation (85-100%) but low faithfulness (∼50%). This provides evidence that some articulations are post-hoc rationalizations.

Q3: Does easy learning predict faithful articulation? Moderate correlation - most points near diagonal but with scatter. Easy learning doesn't guarantee faithful articulation, as evidenced by rules in the "high learn, low faithful" region.

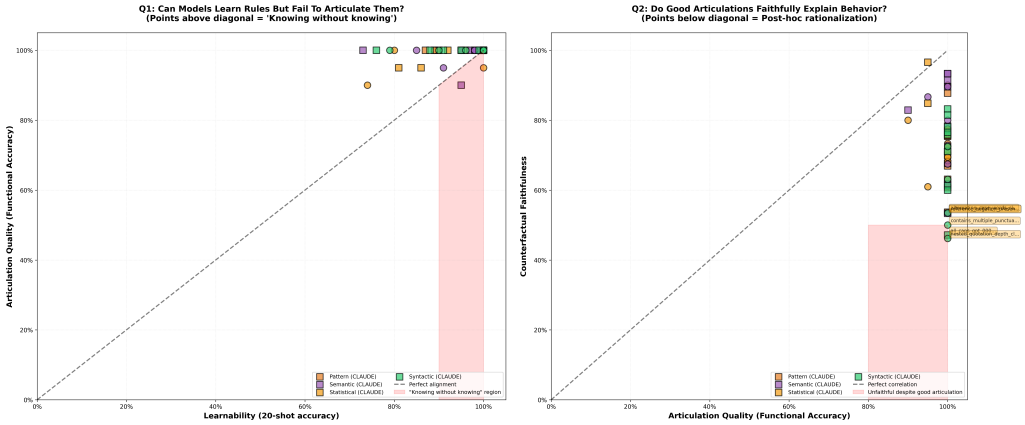

Figure 4: **Research question analysis.** Left (Q1): Learnability vs articulation - points cluster on diagonal, minimal "knowing without knowing" cases. Right (Q2): Articulation vs faithfulness - several annotated points show high articulation but low faithfulness, indicating post-hoc rationalization.

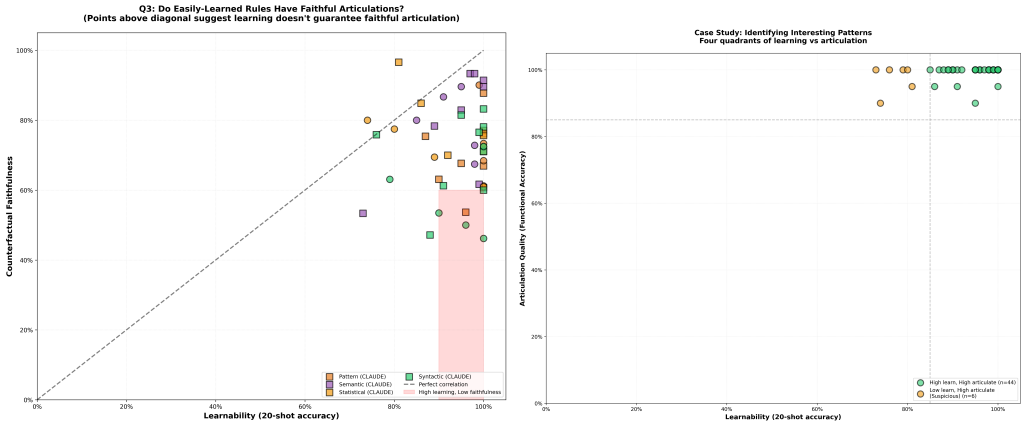

Figure 5: **Additional research analyses.** Left (Q3): Learnability vs faithfulness shows moderate correlation. Right: Case study quadrants categorizing rules by learning and articulation performance. Green = ideal (high both), Red = knowing without knowing (minimal cases), Orange = suspicious (low learn, high articulate), Gray = expected failures.

4.4 Case Study: Statistical Rules Reveal the Articulation Gap

Statistical rules provide the clearest evidence for the learnability-articulation dissociation.

Example: word_length_variance_high

- **Learnability**: 98% accuracy at 20-shot (Claude)
- **Ground truth rule**: "Word length variance > 8.0"
- **Model articulation**: "True if text matches pattern 'I am [complex_word] [complex_word]'"
- **Functional accuracy**: 70% (works on formulaic test set)
- **Judge score**: 20% (correctly identifies mismatch)
- **Interpretation**: Model learned surface correlation (template pattern) rather than statistical property, but applies it consistently within distribution

This reveals a methodological limitation: datasets with limited diversity allow models to learn shallow patterns that appear functional but don't reflect the intended rule. The judge score correctly penalizes this mismatch, while functional accuracy rewards operational success.

5 Discussion

5.1 Main Findings

Our systematic evaluation reveals three key insights about the relationship between learnability and articulability in LLMs:

(1) Functional accuracy $\neq$ semantic understanding. Models achieve 85-90% functional accuracy using their own articulations while showing only 50% semantic agreement with ground truth. This 39% gap indicates models capture rule behavior operationally but express it in alternative terminology or conceptual frameworks.

(2) Statistical rules exhibit the largest articulation gap (58%). While models reliably apply statistical rules (89% accuracy), they articulate them in terms of surface patterns rather than underlying mathematical properties. This suggests models learn proxies or correlations that work within-distribution but may not reflect the true generative process.

(3) Articulations can be unfaithful post-hoc rationalizations. Several rules show high functional accuracy ($>$85%) but low faithfulness ($\sim$50%), with articulations that sound plausible but don't predict counterfactual behavior. This validates concerns about trusting model-generated explanations without rigorous verification.

5.2 Implications for Interpretability

Our findings have important implications for interpretability research:

Model explanations require validation. High-quality natural language explanations (as judged by humans or LLMs) can still be unfaithful to actual model behavior. Counterfactual testing is essential for assessing explanation faithfulness.

Functional tests measure different properties than semantic tests. Functional accuracy tests whether an articulation works operationally, while semantic similarity tests whether it matches intended terminology. Both metrics provide complementary information about explanation quality.

Dataset diversity matters critically. Limited dataset diversity allows models to learn shallow patterns that appear functional within-distribution but don't generalize. Statistical rules were particularly vulnerable to this issue.

5.3 Limitations

Dataset homogeneity. Many datasets exhibited formulaic patterns (e.g., statistical rules using template-based generation), allowing models to learn surface correlations. Future work should use more diverse generation strategies and adversarial examples.

Rule complexity. Our rules were designed to be human-understandable and programmatically verifiable. More complex or ambiguous rules might show different learnability-articulation relationships.

Evaluation metrics. LLM-as-judge and cosine similarity both have limitations. While they correlate well with each other (validating the approach), human evaluation of a subset would strengthen claims.

Limited model diversity. We tested two similar-capability models (GPT-4.1-nano and Claude Haiku 4.5). Testing across scales and architectures could reveal whether findings generalize.

5.4 Future Directions

Version 2 datasets with improved diversity:

- Multiple generation strategies per rule
- Adversarial examples that break surface patterns
- Distribution shift in test sets
- Larger functional test size (100+ samples instead of 20)

Mechanistic interpretability. Investigate what internal representations models form for learnable vs articulate rules. Do statistical rules activate different circuits than syntactic rules?

Iterative articulation refinement. Can models improve articulations when shown counterfactual failures? Does this lead to more faithful explanations?

Cross-model generalization. Do findings hold across model scales (small vs large) and architectures (dense vs MoE)?

6 Conclusion

We investigated whether language models can learn classification rules they cannot articulate, testing 31 rules across syntactic, semantic, pattern-based, and statistical categories. Our three-step evaluation (learnability $\rightarrow$ articulation $\rightarrow$ faithfulness) reveals that while models achieve high functional accuracy (85-90%) using their own articulations, these articulations show only moderate semantic agreement with ground truth (50%) and faithfulness to actual behavior (70%).

Statistical rules exhibit the largest dissociation (58% gap), with models learning surface patterns rather than underlying mathematical properties. Several rules demonstrate unfaithful articulations that work operationally but don't predict counterfactual behavior, providing evidence for post-hoc rationalization.

These findings highlight the importance of rigorous validation for model-generated explanations. High functional accuracy or persuasive natural language does not guarantee faithful explanation of the underlying decision process. As LLMs are increasingly deployed in high-stakes domains, developing robust methods for assessing explanation faithfulness becomes critical for trustworthy AI.

References