MERIT: Mechanistic Explainability of Reasoning Integrity and Transparency

Richard Shan
North Carolina School of Science and Math
Department of Data Science
August 2025

Abstract

Large language models (LLMs) are increasingly trusted in domains such as education, research, and decision making. Despite their widespread adoption, fundamental questions remain about alignment and logic in LLM reasoning. This assurance gap poses significant risks in high-stakes domains where reasoning integrity is critical. This study develops a mechanistic framework for reasoning verification by examining internal model computations rather than relying solely on behavioral outputs. We analyzed DeepSeek-R1 Distill Llama-8B latent feature activations across 2,000 mathematical problems of varying difficulty and domain. Employing Sparse Autoencoders for feature extraction, we identify specific internal features directly tied to reasoning behavior. We develop six novel metrics to quantify reasoning quality by logical structure, self-corrections, and coherence. Through correlation and causal intervention analysis across difficulty levels, we demonstrate that general reasoning features fracture into domain-specific specialists under increased task complexity. Experimental manipulation of identified features produces significant causal effects on reasoning behavior, establishing that internal features functionally control specific aspects of reasoning quality. At higher mathematical difficulties, domain-experts emerge for geometry, number theory, and other mathematical subdomains. Additionally, we demonstrate the existence of distinct feature-governed reasoning modalities: a concise calculation-oriented mode and a verbose explanation-oriented mode We establish that LLM reasoning is predictable by its internal feature structure, which is difficulty-dependent and domain-specialized. This framework enables assurance of reasoning models based on legitimate internal monitoring, rather than output correctness alone, with implications for all reasoning-enabled AI applications.

1 Introduction

1.1 Justification

Large language models (LLMs) have become central to the modern landscape of artificial intelligence [1]. Their capacity to generate detailed solutions and solve problems has already transformed how students learn, how researchers work, and how organizations make decisions. Models such as GPT, Claude, and DeepSeek are now used in tutoring platforms, data analysis systems, and technical support workflows [2]. Increasingly, these systems are being deployed as *reasoning engines*: tools expected not only to provide answers but also to justify them [3].

Reasoning, however, is the least transparent and most fragile capability of LLMs [4]. While they can generate chain-of-thought explanations, these traces do not necessarily reflect authentic internal processes [5]. In many cases, LLMs even fabricate reasoning steps that appear plausible but are not indicative of the computations used in producing the answer [6] [7]. They may disregard instructions to include or exclude reasoning, generating rationales inconsistently. When optimized to please human evaluators, they may output "reasoning-like" text because it maximizes reward, not because it corresponds to actual inference [8] [9]. These behaviors reveal a deep problem: correct answers and even coherent explanations do not guarantee trustworthy reasoning [10] [11]. Reasoning thus cannot be understood from a single benchmark alone [12]. Easier datasets may make reasoning appear generic, while more challenging datasets reveal whether reasoning features hold or fracture.

The risks of this disconnect are substantial. In education, learners may adopt fabricated reasoning as valid [13]. In science and engineering, spurious justifications may obscure flawed conclusions. In policy, reasoning traces shaped to persuade rather than to reason may introduce systematic bias [14]. As LLMs become integrated into high-stakes domains, trust cannot rest solely on correctness of outputs [15]. Assurance requires evidence that genuine reasoning processes are occurring inside the model [16].

This study aims to provide such evidence. I develop a methodology to move beyond binary correctness to examine mechanistic reasoning assurance. By analyzing the internal representations of a reasoning-optimized LLM, I examine whether reasoning features can be isolated, characterized, and connected to reasoning quality. I further compare LLM activations and feature interpretations across datasets of increasing difficulty and identify reasoning feature subdomain specialization.

1.2 Mechanistic Interpretability

Each recent frontier model release has focused on optimizing LLM performance on specific performance benchmarks. While these models represent significant progress, little work has been done in developing interpretability techniques. Recent advancements in the field of mechanistic interpretability have begun to make it possible to study the internal activations of LLMs at a finer scale [17]. Sparse autoencoders (SAEs) have emerged as a promising tool for this task [18] [19]. An SAE compresses the high-dimensional activations of a model layer into a set of sparse features that each activate selectively on particular inputs. The sparsity constraint encourages specialization: instead of every feature activating weakly all the time, many features activate strongly only in specific contexts. This structure allows researchers to identify features that align with recognizable concepts [20]. For instance, SAEs have been used to reveal features that track when text is written in Python code, or that detect stylistic markers such as formality [21]. In effect, SAEs allow researchers to investigate an LLM and identify recurring, interpretable patterns in its activations. While prior work has focused on topical or stylistic features, their potential to capture reasoning processes remains largely unexplored [22].

1.3 Reasoning Evaluation

While mechanistic interpretability has been applied to identify knowledge and style features, reasoning has received far less attention. Reasoning involves multi-step inference, abstract representations, and intermediate calculations. Whether such processes can be captured as features in LLMs, and whether these features are shared across domains or specialized within them, remains an open problem [22].

In parallel, evaluation of reasoning in LLMs has developed a different set of tools. The existing GSM8K benchmark, for instance, is a dataset that tests whether models can solve reasoning-based problems at a grade-school difficulty [23]. Higher complexity datasets, such as the Olympiad mathematics dataset, extend evaluation to more difficult questions via adding domains of number theory, geometry, and combinatorics [24]. These benchmarks have become central to measuring progress in reasoning ability. However, these benchmarks almost exclusively measure *correctness*, or if the expected dataset answer exactly matches the LLM output. They ask whether the model ultimately arrives at the right answer, with little regard to

intermediate reasoning steps. Even when chain-of-thought explanations are collected, there is little effort on examining whether those explanations correspond to genuine internal processes. A model can achieve high accuracy by recalling memorized patterns, exploiting superficial cues, or fabricating plausible reasoning [10], none of which constitute true reasoning.

This focus on correctness leaves reasoning assurance underdeveloped. Knowing that a model can answer a question correctly does not tell us how it reached that answer, nor whether that same reasoning can generalize to a new question or domain. Assurance requires evidence that identifiable reasoning processes do occur inside the model, and that these processes are linked to reasoning quality.

Critically, most existing evaluations study models on a single benchmark. This practice obscures an important dimension of reasoning: its sensitivity to problem difficulty. On GSM8K, reasoning features may appear to generalize broadly. Yet when the same features are examined on Olympiad problems, they may instead fail to reproduce reasoning and only demonstrate pattern-matching capabilities. Alternatively, increasing difficulty can fracture features into domain-specific patterns, potentially revealing modular structure invisible on simpler tasks [25]. Without contrasting behavior across difficulty levels, evaluations risk misclassifying features as general when they are in fact specialized.

There is a need for more work at the intersection of interpretability and reasoning evaluation. Interpretability demonstrates that topical features exist within LLMs, while benchmarks demonstrate that reasoning behavior emerges. However, the validation of reasoning via internal evidence remains missing.

This study addresses that gap by developing a methodology for mechanistic reasoning assurance. I examine reasoning mechanisms within DeepSeek-R1 Distill Llama-8B, a representative reasoning-optimized LLM. The approach uses sparse autoencoders and judge-LLM domain classification to classify the roles and domains of extracted features. I introduce six new reasoning-quality metrics designed to quantify the structure and reliability of model outputs beyond correctness alone. This study applies the methodology across GSM8K and Olympiad mathematics to test how reasoning features that appear generic on simpler problems fracture into domain-specific patterns under complex tasks. Together, this methodology combines feature extraction with automated interpretation, introduces new reasoning-quality metrics, and uses dataset difficulty as an experimental variable to probe reasoning assurance mechanistically.

2 Methods

2.1 Data

In this study, the primary data consisted of two complementary sets of reasoning traces generated with DeepSeek R1 Distill Llama-8B: 1,000 LLM responses to problems from grade-school math and 1,000 responses to problems from advanced Olympiad mathematics. For each problem and response instance, we captured LLM external and internal signals of reasoning. The external layer included the model's step-by-step reasoning output, final answer, and process metadata such as response length, token counts, and formatting structure. The internal layer consisted of residual stream activations, and a signature of the top 50 most active features for that problem. Because the same LLM and autoencoder were used across both datasets, feature identifiers indices are consistent, enabling direct cross-dataset comparison. Finally, every reasoning trace was evaluated using six reasoning-quality metrics (detailed in Section 2.5), providing a unified framework to analyze outputs, activations, and the impact of task difficulty.

2.1.1 Model Selection

This study examined the internal reasoning processes of a frontier reasoning-tuned large language model (LLM). All experiments were conducted using DeepSeek R1 Distill Llama-8B, a distilled variant of the DeepSeek R1 reasoning model built on the LLaMA-3.1-8B transformer architecture. DeepSeek R1 Distill belongs to a recent class of reasoning-optimized LLMs, which are trained specifically to output multi-step chain-of-thought style explanations.

The DeepSeek R1 family was developed with training on chain-of-thought data and reinforcement learning objectives to encourage multi-step inference. The LLaMA-distilled variant used in this study was produced via knowledge distillation, in which a student model is trained to mimic the behavior of a larger teacher model. In this case, the teacher was DeepSeek R1, and the student was optimized to capture reasoning traces, allowing reasoning ability itself to be transferred during distillation. This property made DeepSeek R1 Distill particularly suitable for an extended mechanistic study requiring thousands of forward passes. All model loading and inference were conducted through the HuggingFace Transformers library.

Internally, LLaMA-family models follow the transformer architecture, in which each block consists of attention and feedforward sublayers connected by a residual stream. The residual stream can be understood as the model's "running notebook". At each block, it accumulates contributions from prior layers and passes them forward. More formally, the residual stream is the sum of the attention and multi-layer perceptron outputs at each block, making it the locus for representational information that drives the model's next steps. Hence, sparse autoencoders focused on the residual stream can potentially recover features aligned with human-interpretable concepts. Probing this stream offers direct access to the latents that the model actively maintains for reasoning, before they are collapsed into logits for output.

In this study, activations were collected from the residual stream of block 19 (blocks.19.hook_resid_post), a mid-to-late layer out of Deepseek's 32 layers. Early layers primarily encode token-level embeddings, while final layers reflect surface-form outputs. By contrast, mid-level residual streams contain semantically rich intermediate representations thought to support reasoning. These activations were subsequently analyzed with a Sparse Autoencoder (SAE) trained for this model family, yielding compressed, interpretable feature vectors for each problem instance, discussed further in 2.2.

2.1.2 Dataset Selection

Two datasets of contrasting difficulty were used to probe reasoning features. The Grade-School Math 8K dataset (GSM8K) is a benchmark of arithmetic and word-based problems designed for grade-school level reasoning [23]. Problems typically require a handful of explicit steps and involve straightforward operations such as ratios, addition, subtraction, and basic algebra. A subset of 1,000 problems was randomly sampled for analysis. The Olympiad Mathematics (Olympiads) dataset is a subset of 1,000 competition-level problems drawn from sources such as AMC, AIME, USAMO, and IMO archives [24]. These problems span advanced algebra, number theory, geometry, and abstract math. Unlike GSM8K, Olympiad tasks often demand symbolic manipulation, multiple layers of inference, or proof-oriented reasoning.

The use of these two datasets enabled a comparative difficulty design: GSM8K served as a baseline where reasoning features may appear broad and generic, while Olympiad problems acted as a stress test, revealing whether those same features fractured into specialized subdomains under higher complexity.

2.1.3 Problem Domain Classification

As mentioned in section 2.1.2, I pull problems from two datasets of varying difficulties. To investigate LLM reasoning feature specialization, it is critical to classify these problems into their respective mathematical subdomains. We classify problems into one of five domains: basic algebra/arithmetic, geometry, advanced algebra, number theory & combinatorics, and abstract math/analysis. I use Llama 4 Maverick as a "judge" LLM and provide it with example problems and a specialized prompt for classification of problems. The judge LLM is used to automatically classify each problem into one of those subdomains, which is important for later analysis.

2.1.4 Reasoning Metrics

For each problem instance, the model was prompted to produce a step-by-step reasoning trace along with its final answer. In addition to the output, we record the metadata of response length, line breaks, formatting markers, and the feature activations. We record the indices of the 50 highest activating features on each problem, providing that they are above our minimum activation threshold of 50. Feature indices are constant across all problems and datasets, as they are unique to the Deepseek model itself. Thus, we can analyze feature performance across problems and domains given their activation data.

Each response was further evaluated against a set of six reasoning-quality metrics: step indicators, calculations, reasoning words, explanation phrases, corrections, and structured indicators. Their formal definitions and implementations are described in Section 2.5.

2.2 Sparse Autoencoder Feature Extraction

2.2.1 Motivation for SAE

The internal states of a large language model are high-dimensional and densely entangled, making them extremely difficult to interpret directly. In DeepSeek R1 Distill Llama-8B, the residual stream at block 19 has 4,096 dimensions, yet the model must represent far more distinct concepts than this dimensionality allows. As a result, many features are superposed: multiple unrelated concepts overlap in the same neurons, and conversely, individual concepts are spread across many neurons. Superposition is an efficient representational strategy, but it renders raw activations nearly opaque, since no single value or direction corresponds cleanly to a recognizable idea.

Sparse autoencoders (SAEs) are designed to address this problem. By training an overcomplete dictionary of features and enforcing sparsity in the hidden layer, SAEs map dense activations into a higher-dimensional space where only a small fraction of features fire at once. This sparsity forces specialization, encouraging different features to encode distinct, consistent patterns rather than overlapping mixtures. In other words, the autoencoder "untangles" superposed signals, yielding features that can be studied as discrete representational units.

Empirical work has shown that SAE features often correspond to semantically meaningful concepts, such as detecting when text is Python code, identifying stylistic formality, or tracking arithmetic operators. This suggests that even abstract processes like reasoning may be recoverable if they are encoded in a superposed form. For reasoning assurance, this capability is critical: if reasoning structures exist in the residual stream but are hidden in tangled activations, SAEs provide the tool to expose them as interpretable features.

2.2.2 Architecture of the SAE

A sparse autoencoder (SAE) is a neural network designed to map dense residual activations into a higher-dimensional but sparse latent space, where only a small subset of features fire for any given input. This architecture is motivated by the superposition hypothesis: in a transformer residual stream, the number of features the model must represent greatly exceeds the dimensionality of the hidden state, forcing many unrelated concepts to overlap in the same neurons. SAEs address this by learning an overcomplete dictionary of features, combined with sparsity constraints that encourage disentanglement.

Let $x \in R^d$ denote a residual stream activation vector, with $d = 4096$ for block 19 in DeepSeek R1 Distill Llama-8B. The encoder projects x into a higher-dimensional hidden vector $h \in R^k$ with $k \gg d$:

$$h = \sigma(W_e x + b_e)$$

where $W_e \in R^{k \times d}$ and $b_e \in R^k$ are encoder parameters, and σ is a rectified linear unit (ReLU) to ensure nonnegative activations. The decoder reconstructs the input as: $\hat{x} = W_d h + b_d$, with decoder parameters $W_d \in R^{d \times k}$, $b_d \in R^d$.

Further, the autoencoder is trained to minimize a reconstruction loss with an additional sparsity penalty:

$$\mathcal{L} = |x - \hat{x}|_2^2 + \lambda |h|_1$$

The L_1 term ensures that only a small subset of hidden units fire strongly for any given input. In practice, some SAEs also use top-k sparsity, in which only the k largest hidden activations are retained and the rest are zeroed out. Both approaches constrain overlap, reduce interference, and push different features to specialize. This is the mechanism by which SAEs "untangle" superposed signals into more monosemantic features.

Our SAE architecture, illustrated in Figure 1 below, expanded each 4,096-dimensional activation into ~16,000 hidden units. Despite this larger space, sparsity meant that typically fewer than 50 features fired per problem instance. This design combines a large dictionary of possible features with only a few active features per input, making the resulting features both expressive and interpretable.

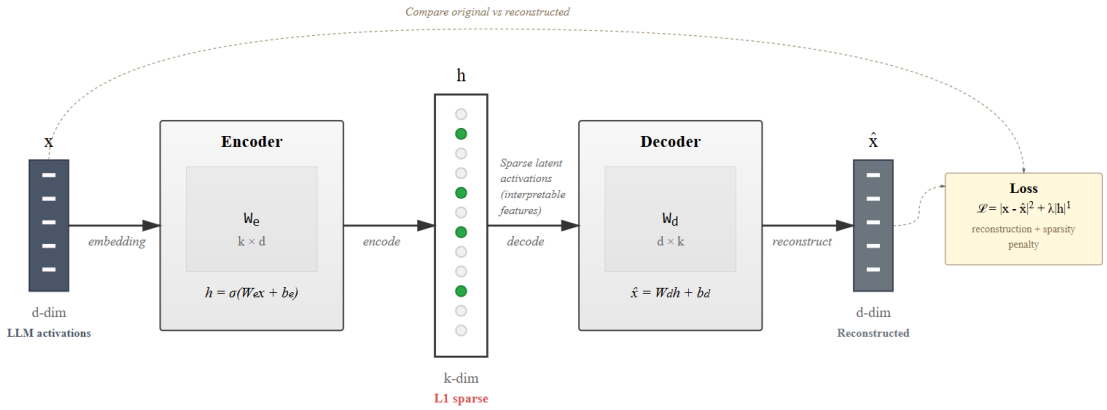

Figure 1*:* Sparse autoencoder workflow for mapping LLM activations into sparse features and reconstructing inputs. Figure created by student researcher using React in 2025.

Each hidden unit corresponds to a feature direction in activation space: the encoder weights $W_e[j]$ define when feature j fires, and the decoder weights $W_d[:, j]$ describe how it contributes back to reconstruction. A feature is considered interpretable if it activates consistently in the presence of a coherent signal (e.g., addition operations, formal proof language). Prior work shows that SAEs trained on LLM activations reliably recover such monosemantic units.

For each problem instance in GSM8K and Olympiad datasets, residual activations at block 19 were passed through the pretrained SAE. The top-50 most active features above a trivial

threshold were retained, along with their normalized activation magnitudes, forming a feature signature of the model's internal reasoning state. These signatures served as the basis for automated interpretation and comparative analyses.

2.2.3 Hook Point Selection

This study probes activations from the residual stream at block 19 of DeepSeek R1 Distill Llama-8B. The residual stream integrates contributions from both the attention and MLP sublayers, preserving the cumulative state of computation. Block 19 was chosen out of a total of 32 because mid-to-late transformer layers have been shown to hold the richest internal representations. Early layers are dominated by word identity and positional encoding, while the very last layers are tuned to produce output tokens. By contrast, middle layers contain the intermediate abstractions the model uses to carry out multi-step reasoning. Probing at block 19 therefore provides access to reasoning processes as they unfold, at a point where the computation is neither too shallow nor already collapsed into final predictions.

2.2.4 SAE Configuration

We used a pretrained sparse autoencoder (SAE) distributed as deepseek-r1-distill-llama-8b-qresearch via the sae_lens library and loaded it at the residual-stream hook point block 19. We report the encoder matrix dimensionality $W_{\text{enc}} \in R^{k \times d}$ from the SAE encoder weight matrix at runtime with $d = 4096$.

For each problem instance, we extracted hidden states from layer index 19 and mean-pooled across the sequence dimension to obtain a single residual vector per prompt before SAE encoding. The resulting latent vector was sparse; we retain only the top-50 features per instance, with activations below 0.01 discarded. The indices and magnitudes of these active features constituted the feature signature used in all subsequent analyses.

2.2.5 Feature Example

To provide a concrete illustration of the type of units recovered by the sparse autoencoder, we highlight Feature F25111, a feature that emerged consistently in the grade-school GSM8K dataset. Figure 2 shows the proportion of GSM8K problems in each mathematical domain where F25111 fired above the activation threshold of 0.01. The feature was most active in geometry (44%), with substantial coverage in arithmetic (34%) and

percentages/ratios (32%). It appeared less frequently in algebra (20%), and did not activate for probability/statistics (0%).

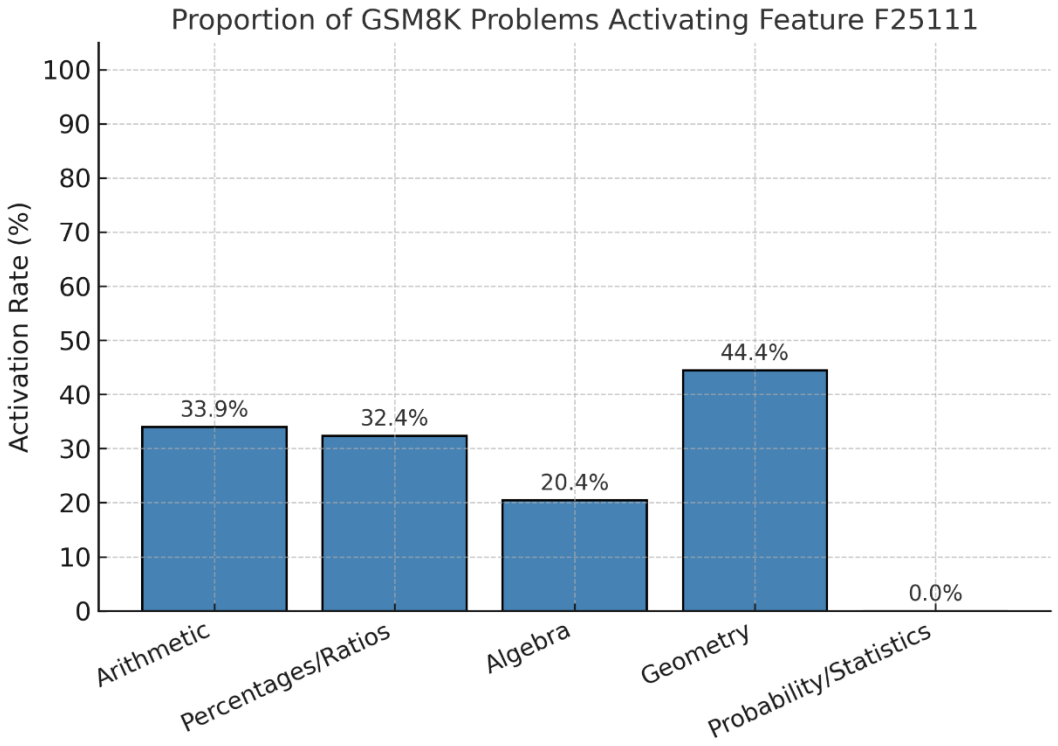

Figure 2: Activations of Feature 25111 across GSM8K mathematical domains. Figure created by student researcher using Python in 2025.

This distribution suggests that F25111 encodes a broad numerical reasoning signal that generalizes across elementary mathematical tasks, with particularly strong alignment to quantitative reasoning. The absence of F25111 activations in probability/statistics indicates that this feature does not generalize across all forms of mathematical reasoning. Instead, it specializes in deterministic, step-based computation and fails to respond when problems involve uncertainty or distributional reasoning.

This example demonstrates interpretability insights gatherable from sparse features. Discrete internal activations can be mapped to human-recognizable concepts. While F25111 appears to capture a generic quantitative reasoning process in GSM8K, later analyses will examine whether such features remain broad under Olympiad-level difficulty or fracture into domain-specialized patterns.

2.3 Automated Feature Interpretation

For each problem instance, the SAE outputs a set of indices corresponding to the most active features and their associated activation magnitudes. However, while sparse features are more coherent than their raw activations, they are not inherently self-describing. Each feature is defined only by its activation pattern, leaving open the question of what reasoning behavior it corresponds to. To address this, I develop an automated sparse feature interpretation pipeline (auto-interpretation) using a judge language model, as illustrated in Figure 3. As each feature has a unique identifier, it is possible to assemble a feature profile based on its activations across many different queries. Given a large amount of feature activation records on various queries, we can identify which queries correlate with activations of specific features. I use Llama 4 Maverick as the judge LLM, with a customized prompt tuned for feature interpretation.

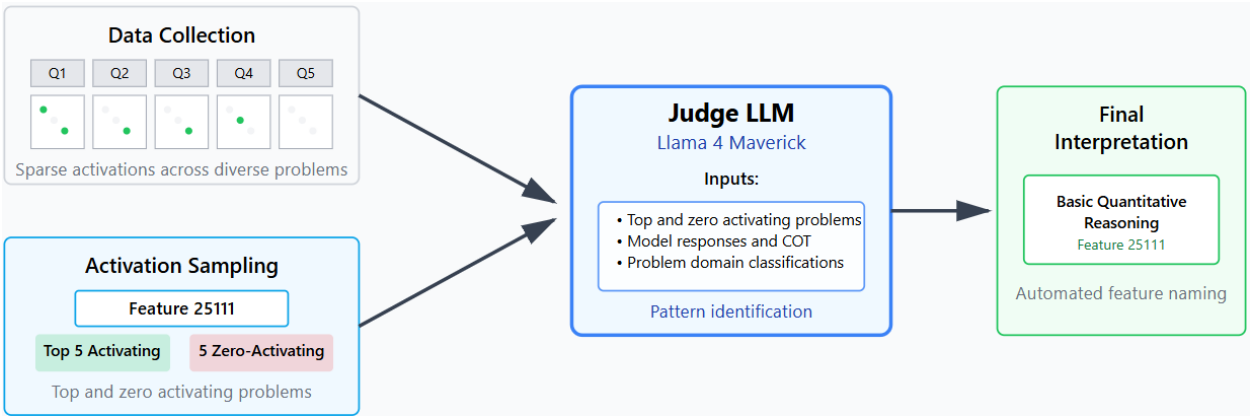

Figure 3: Automated feature interpretation via judge LLM. Figure created by student researcher using React in 2025.

To interpret a particular feature, we first isolate within the dataset the top 5 problems that caused the highest activations of that feature, along with the 5 zero-activating problems. The judge is given the text of the query, the response, the feature activations, as well as the problem classification as outlined in section 2.1.3. The judge model is instructed to identify patterns between activations and docs, and the final resulting interpretation is saved.

2.4 Reasoning Output Generation

Reasoning traces were generated by prompting DeepSeek R1 with natural language questions from each dataset. For GSM8K, 1,000 problems were randomly sampled from both the training and test splits of the *openai/gsm8k* benchmark. For Olympiad mathematics, 1,000 problems were sampled from the *aslawliet/olympiads* dataset, which aggregates from the AMC,

AIME, and IMO archives. Each problem was transformed into a standardized prompt of the form:

Question: < problem text >

Answer:

The model was then allowed to autoregressively generate a continuation under constrained sampling. For GSM8K, generation was capped at 300 new tokens to match the short length of arithmetic reasoning traces. After finding token cutoffs during testing, the cap was increased to 500 new tokens for Olympiad problems which allow for extended proof-like arguments and longer symbolic derivations. A temperature of 0.3 was used to reduce stochastic variation while preserving the diversity inherent to reasoning.

For each generation, we systemically log as metadata the raw response text, response length, and context length alongside metrics such as generation time and length. We classify each generation into their mathematical subdomain. For both datasets, six core metrics were extracted: step indicators, calculations, reasoning words, explanation phrases, corrections, and structural markers (Section 2.5). For Olympiad data, five further proof-oriented metrics were applied, including proof language, logical connectives, mathematical formalism, case analysis, and generalization attempts.

To account for potential failure cases, each problem was also scored on five complexity dimensions by an auxiliary LLM-judge (computational load, reasoning steps, conceptual abstraction, setup complexity, and mathematical domain). Where API failures or parsing errors occurred, a rule-based fallback computed these scores directly from problem text. Fallback usage was logged per instance to maintain dataset integrity.

2.5 Reasoning Quality Metrics

Traditional reasoning benchmarks evaluate models almost exclusively on correctness: whether the final answer matches the ground truth. While useful, this metric alone cannot distinguish between reliable reasoning and spurious success. A model may guess correctly, recall a memorized template, or fabricate a plausible chain-of-thought without engaging in reasoning. To probe reasoning assurance, this study introduced a set of six reasoning-quality metrics, illustrated in Table 1, that quantify the structure and reliability of reasoning traces independent of correctness.

Table 1: Custom reasoning quality metrics and significance. Table created by student researcher in 2025.

Metric	Definition	Example	Significance
Step Indicators	Transitional phrases that mark reasoning progression	“first”, “second”, “third”, “next”, “then”, “finally”	Indicates structured thinking
Calculations	Mathematical operation symbols and computational indicators	“ = ”, “ + ”, “ - ”, “ × ”, “ ÷ ”, “ * ”, “ / ”	Direct measure of quantitative reasoning and computation
Reasoning Words	Logical connectors establishing causal relationships	“therefore”, “so”, “thus”, “because”, “since”, “hence”, “consequently”	Explicit logical reasoning and argument construction
Explanation Phrases	Expressions showing deliberate problem-solving	“let me”, “we need to”, “to find”, “to calculate”, “to solve”, “let’s”	Shows deliberate planning of solution steps
Corrections	Self-identified errors and reasoning revisions	“wait”, “actually”, “correction”, “mistake”, “wrong”	Demonstrates error detection and self-correction
Structured Indicators	Organizational markers for systematic breakdown	“1.”, “2.”, “3.”, “ – ”	Decomposition for complex multi-step reasoning

These six metrics together provide a structured framework for evaluating reasoning quality. By capturing different signals for organization, computation, logical coherence, planning, and self-correction, they allow reasoning traces to be quantified along multiple dimensions rather than treated as binary success or failure. Importantly, each metric isolates a distinct facet of reasoning behavior. We will employ these custom metrics to great effect in later analysis sections.

We apply these custom metrics to a sample GSM8K problem, shown below in Figure 4. The problem involves counting the number of animal legs on a farm. The model’s response includes multiple reasoning behaviors: arithmetic calculations (“10 * 2 = 20”), logical connectives (“so”, “then”, “therefore”), and corrections (“wait, let me think again”). On this question, we detect 2 step indicators, 26 calculations, 25 reasoning words, and 3 explanation phrases, alongside other markers. These statistics provide a process-level fingerprint of the actual reasoning trace.

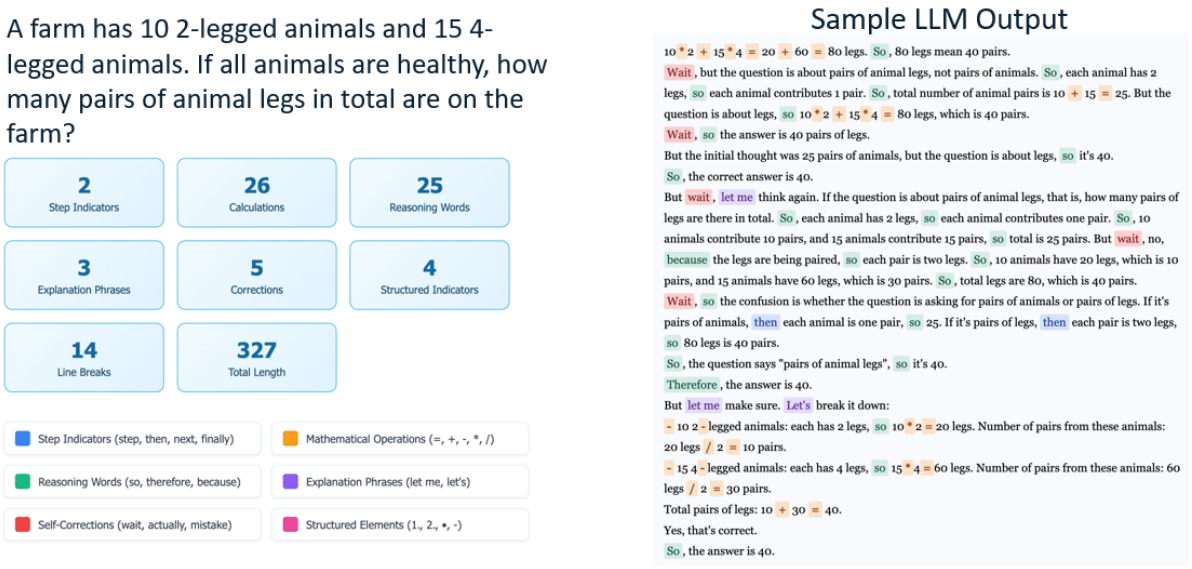

Figure 4: Example application of reasoning-quality metrics to LLM reasoning trace. Figure created by student researcher using React in 2025.

By quantifying these features, reasoning quality can be studied as a continuous variable rather than a binary success/failure. For example, a response that reaches the wrong final answer but demonstrates 20 calculations and consistent step indicators still is reflective of genuine reasoning, whereas a correct answer with no intermediate markers may indicate pattern-matching or guessing.

2.6 Comparative Difficulties

A key objective of this study was to examine how reasoning features behave as task complexity increases. GSM8K problems require only a few arithmetic or algebraic steps, providing a baseline where reasoning features may appear broad and generic. Olympiad tasks, by contrast, demand extended symbolic reasoning and domain-specific skills such as geometry, combinatorics, or number theory, creating a natural stress test for reasoning under complexity. Both datasets were processed through the same pipeline, using identical prompting, inference, feature extraction, and reasoning-quality metrics. This design ensured that any observed differences could be attributed to difficulty and domain rather than inconsistencies in generation or analysis.

3 Results and Discussion

3.1 Extracted Feature Distributions

By comparing GSM8K and Olympiad instances, we observe distinct patterns that reflect the complexity of reasoning engaged. On GSM8K, the number of active features per problem is consistently low. Most problems triggered only a handful of features, with a median in the single digits. The distribution is tightly concentrated, as shown in Figure 5, indicating that elementary problems are solved with relatively broad and generic reasoning representations. This suggests that the model can reuse a small set of internal features across many simpler tasks without requiring deep specialization.

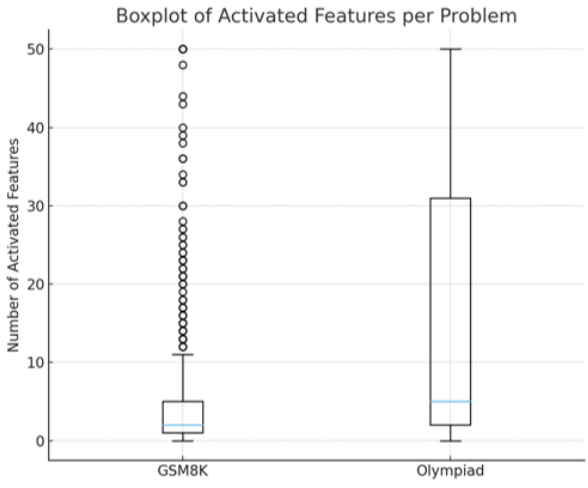

Figure 5A: Boxplot of activated features per problem.

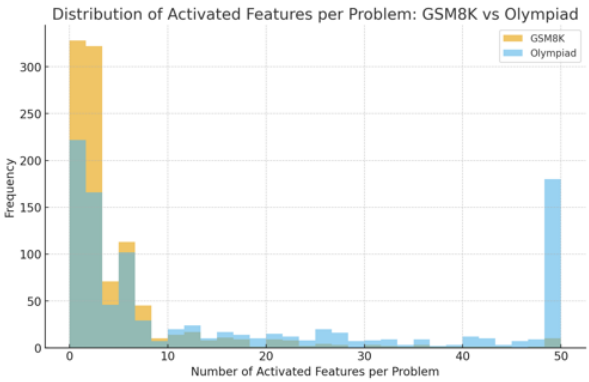

Figure 5B: Distribution of activated features per problem

Figure 5: Comparisons of activated features per problem between datasets. Figure created by student researcher using Python in 2025.

By contrast, Olympiad problems produce a much denser activation landscape. The histogram shows a pronounced shift toward larger numbers of features. During data collection, we cap recorded features at the top 50 highest-activating features. We see that the Olympiad queries nontrivially fire 50+ features above the activation threshold a significant portion of the time. The boxplot comparison in Figure 5 highlights this contrast: Olympiad problems exhibit a higher median, a broader interquartile range, and a longer upper tail.

These patterns suggest that problem complexity directly influences the internal feature budget. Easier tasks can be handled with catch-all generic reasoning features, whereas complex Olympiad-level tasks demand richer and more specialized activation signatures. This evidence aligns with the hypothesis that reasoning under difficulty is not just longer or noisier, but structurally different in how it engages the model's internal representational space.

3.2 Reasoning Feature Testing

For all subsequent results, the central focus is on features that genuinely capture reasoning rather than incidental patterns. Although the dataset was curated to emphasize reasoning, the top 50 activating features recorded per problem inevitably include units that respond to superficial cues such as formatting or domain-specific vocabulary. Distinguishing these from features that drive structured inference is therefore critical. To make this separation, I systematically test the correlation between feature activations and the reasoning-quality metrics.

For an example test, recall Feature F25111 which was auto-interpreted in Section 2.3 as a "basic quantitative reasoning" feature. However, to rigorously evaluate F25111's influence on reasoning, I statistically test the relationship of F25111's activation levels in the GSM8K dataset on the step indicator metric as described earlier. Because both variables are continuous, a Pearson correlation test is employed to assess the linear association between feature activation and reasoning quality.

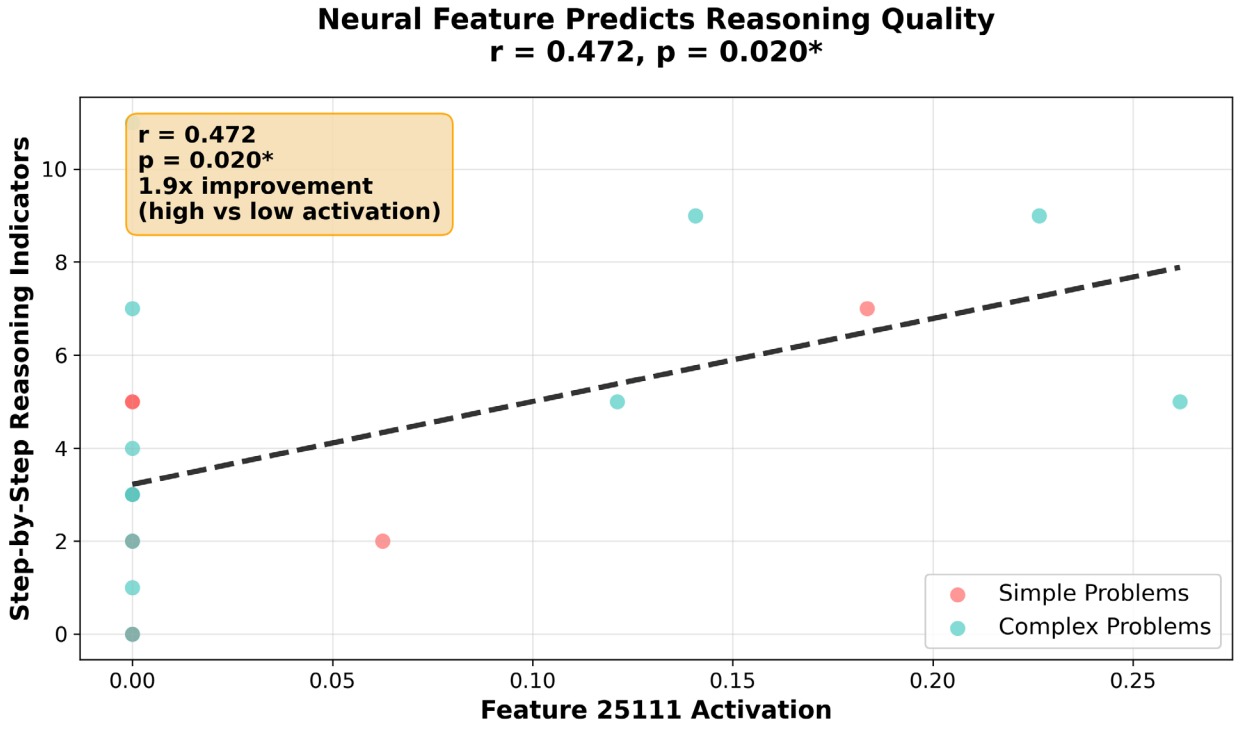

Figure 6: Correlation between Feature 25111 activation and reasoning quality on GSM8K. Figure created by student researcher using Python in 2025.

We correlate F25111's activation magnitude with the step-indicator metric. The result, shown above in Figure 6, reveals a statistically significant relationship that higher activation of

F25111 predicts greater reasoning. Strong activations of F25111 displayed a 1.9x increase in reasoning steps relative to low-activation cases, with a significant p-value of 0.02.

Importantly, this analysis includes zero-activating examples where the feature did not fire at all, indicated by the points with x-coordinate value of zero in the same Figure 6. Excluding them would bias the sample toward successful activations and thus inflate the correlation. By retaining zero-activation data points, we confirm that reasoning quality systematically differs between problems that engage the feature versus those that do not. This strengthens the interpretation that F25111 acts as a genuine feature of structured reasoning. This establishes a foundation for examining whether such reasoning features remain stable across domains, or fracture under higher levels of difficulty.

To test whether feature-reasoning correlations reflect causal relationships, we performed activation interventions on Feature 25111. We systematically manipulated Feature 25111 across three conditions: natural (baseline), suppressed (activation=0.0), and enhanced (activation=2.0). Intervention results shown below in Figure 7 revealed a significant relationship with correction phrases, the number of times the LLM revisited and revised its own reasoning.

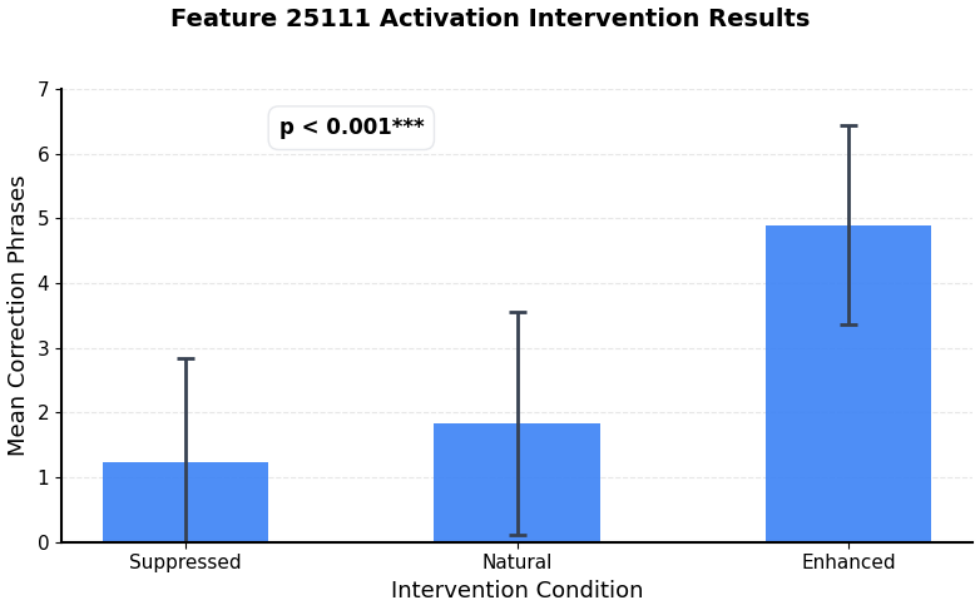

Figure 7: Feature 25111 activation intervention results. Figure created by student researcher using React in 2025.

Under the suppressed condition, when F25111 was silenced, the model produced a mean of only 1.24 corrections per problem. The natural baseline condition yielded 1.84 corrections, representing a 48% increase. Most strikingly, the enhanced condition elicited 4.90 corrections per problem, a 166% increase over baseline and a nearly fourfold increase over suppression.

A one-way ANOVA confirmed these differences were highly statistically significant with $F(2, 147) = 72.18$, $p < 0.001$. The effect size was substantial, with $\eta^2 = 0.495$, indicating that the experimental manipulation of F25111 explained 49.5% of the variance in self-correction behavior. This large effect size demonstrates that F25111 is not merely correlated with reasoning quality but causally drives the model's tendency to monitor and correct its reasoning.

These findings provide strong mechanistic evidence that F25111 functions as an internal reasoning controller. When suppressed, the model generates answers with minimal self-evaluation. When enhanced, it engages in significantly more metacognitive behavior, repeatedly checking and revising its reasoning chain. This pattern aligns with the feature's interpretation as a quantitative reasoning unit: problems requiring numerical inference benefit from iterative verification, and amplifying F25111 intensifies this self-monitoring process.

Critically, this intervention confirms that the correlations observed earlier reflect genuine causal structure. We establish that sparse autoencoder features are not merely descriptive but functionally operative components of the model's reasoning machinery.

3.3 Subject Matter Expert Features

We next examine whether certain features specialize in distinct mathematical subdomains. To do so, we grouped Olympiad problems by domain (basic algebra, geometry, advanced algebra, number theory, abstract mathematics) as described in Section 2.1.3, and measured average feature activation within each domain subset of problems. This approach enables the identification of subject matter expert (SME) features that consistently activate more strongly in one domain relative to others.

Feature	Basic Algebra	Geometry	Advanced Algebra	Number Theory	Abstract Math	Specializes In
Feature 25111	0.250	0.038	0.015	0.028	0.012	Basic Algebra
Feature 53164	0.069	0.352	0.045	0.032	0.018	Geometry
Feature 8690	0.094	0.442	0.082	0.158	0.035	Geometry
Feature 14998	0.403	0.426	0.417	0.832	0.245	Number Theory
Feature 44576	0.460	0.708	0.811	1.305	0.289	Number Theory
Feature 2308	0.056	0.195	0.034	0.018	0.008	Geometry

Figure 8: Specialization of selected features across mathematical subdomains in Olympiad dataset. Figure created by student researcher using React in 2025.

We examine several features of interest, whose domain-specific activations are illustrated in Figure 8. Higher activations in a category are red, and each feature's highest specializing category is boxed. For instance, F25111 highly activates for basic algebra questions, aligning with its earlier interpretation as a quantitative reasoning feature for basic arithmetic. It demonstrates a 6.5x higher activation in the basic algebra domain compared to other domains. Similarly, F53164 and F8690 consistently specialize in geometry, both activating 5x more than other domains. In contrast, F14998 and F44576 show selectivity for number theory.

To rigorously assess whether the observed domain-specific activation patterns were statistically meaningful, we applied a one-way analysis of variance (ANOVA) and the non-parametric Kruskal-Wallis test. The ANOVA evaluates whether mean activation levels differ significantly across the five mathematical domains, under the assumption of approximate normality, while the Kruskal-Wallis test relaxes these assumptions by comparing ranked distributions. Employing both tests ensures robustness, as feature activations are sparse and often skewed. Importantly, zero-activation cases were retained in each group to reflect the full activation profile of a feature, since absence of firing is as informative for specialization as consistent activation.

The results presented in Table 2 provide clear evidence of subdomain selectivity for several features. Feature 8690 exhibited the strongest statistical evidence of specialization. F14998 and F44576 also displayed significant specialization with number theory. F25111 showed weaker dependence, reaching significance under the Kruskal-Wallis test but not under ANOVA, consistent with its role as a broad quantitative reasoning marker rather than a sharply localized expert feature. By contrast, F53164 and F2308 did not exhibit statistically significant differences across domains ($p > 0.05$).

Table 2: Significance testing of feature domain specialization. Table created by student researcher in 2025.

Feature	Domain Specialty	ANOVA p-value	Kruskal-Wallis p-value
25111	Basic Algebra	0.117	**0.031 ***
53164	Geometry	0.287	0.281
8690	Geometry	**0.0018 ***	**< 0.0001 ***
14998	Number Theory	**0.011 ***	**0.004 ***
44576	Number Theory	**0.048 ***	**0.044 ***
2308	Geometry	0.395	0.502

These findings carry several important implications. First, they provide rigorous confirmation that reasoning features are not uniformly distributed across tasks, but instead cluster into domain-specialized units under mathematical reasoning. This establishes mechanistic evidence for a form of modularity: certain internal features act analogously to human subject matter experts, activating selectively when problems fall within their area of competence.

Second, the mixture of significant and non-significant results indicates that the feature space is heterogeneous, containing both general features that operate broadly across problem types and also specialist features that fire predominantly within a single domain. This dual structure suggests that large language models do not rely solely on generic reasoning heuristics, nor do they fully compartmentalize reasoning into isolated silos. Instead, they appear to combine transferable reasoning markers with SME-like components, producing a hybrid representational strategy.

Finally, the emergence of SME features under Olympiad-level tasks supports the claim that increasing difficulty drives representational reorganization. Whereas GSM8K problems were handled with a relatively small set of generic reasoning features, complex Olympiad problems elicited specialized circuits that fractured along mathematical subdomains.

3.4 Generic Math Feature

Previously, features were identified that specialize in distinct mathematical subdomains, consistent with the hypothesis of subject matter expert (SME) units. However, not all features fall neatly into this pattern. Some appear to serve a broader role, activating robustly across multiple domains.

We initially suspected that Feature 25111 was such a “generic math” feature, as it consistently appeared in GSM8K and displayed strong correlations with step-indicator reasoning metrics. Yet, closer analysis revealed that its activations are disproportionately concentrated in basic algebra, and statistical testing confirmed that it is better characterized as an elementary quantitative reasoning unit rather than a truly domain-general feature. This finding demonstrates that apparent generality at lower difficulty can collapse into specialization under Olympiad-level reasoning.

By contrast, Feature 44576 emerges as a far stronger candidate for a genuinely general mathematical feature. While earlier we identified its specialization in number theory, it maintains

the highest activation magnitude across all domains, outperforming even domain-specific SME features in their home categories as shown in Figure 9. For example, in geometry and advanced algebra, Feature 44576 surpasses features specialized to those domains. F44576 is the highest activating feature across all five domains of basic algebra, geometry, advanced algebra, number theory, and abstract math.

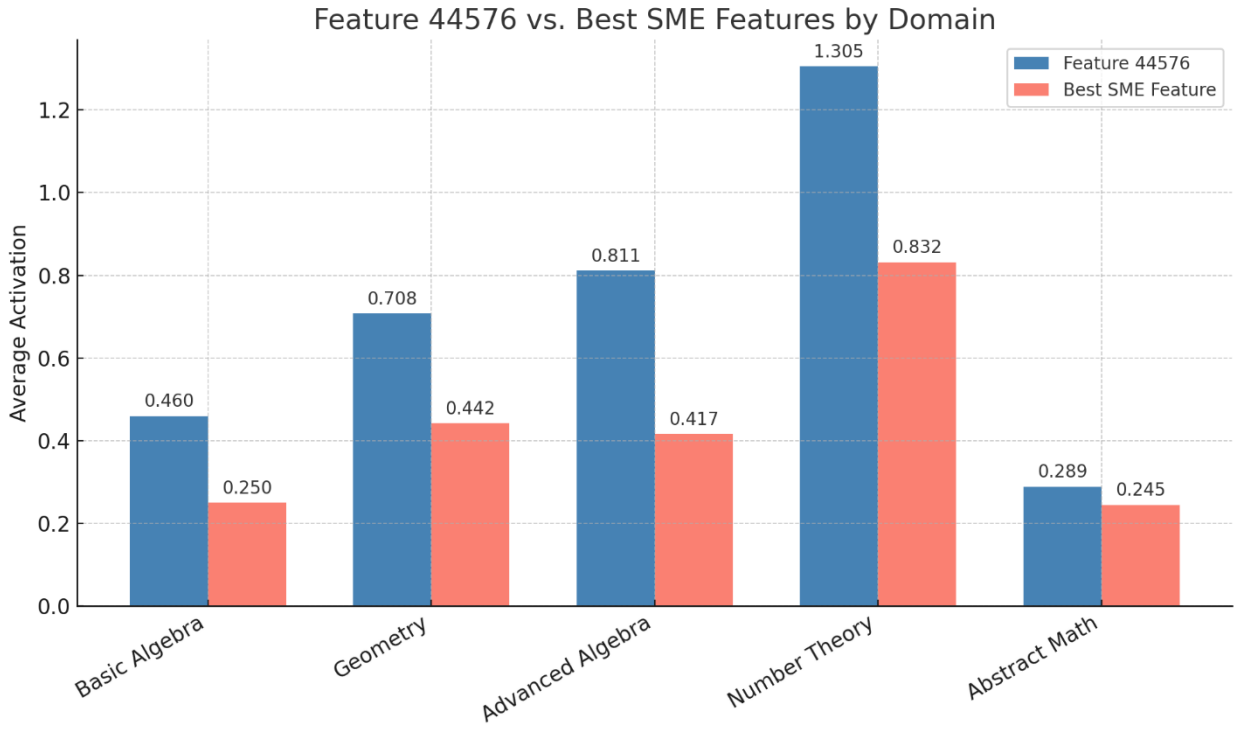

Figure 9: F44576 activations compared to best domain SME features. Figure created by student researcher using Python in 2025.

This pattern may be justified by the longstanding mathematical view that number theory underpins much of mathematics. Mathematicians have maintained that many branches, from algebraic structures to combinatorics and cryptography, ultimately reduce to number-theoretic foundations [27]. If so, then a number-theory aligned feature like F44576 may naturally manifest across diverse domains, serving as a structural generalist rather than a narrow SME.

3.5 Reasoning Modalities

3.5.1 Distinct Modalities

Psychologists and cognitive scientists have long identified unique reasoning modes for humans such as analytical reasoning, moral reasoning, and probabilistic reasoning [28]. Prior work has suggested that LLMs, having been trained on human data, may reason in a similar

fashion [29]. I identify that LLM reasoning in the mathematical domain also occurs in two key modalities: a verbose explanation-heavy teaching mode and a concise calculations-based execution mode.

We first analyze the reasoning quality metrics as described in Section 2.5. We calculate the Pearson correlation of all reasoning metrics over the 1,000 records (n=1000) of the GSM8K data, as shown below in Figure 10.

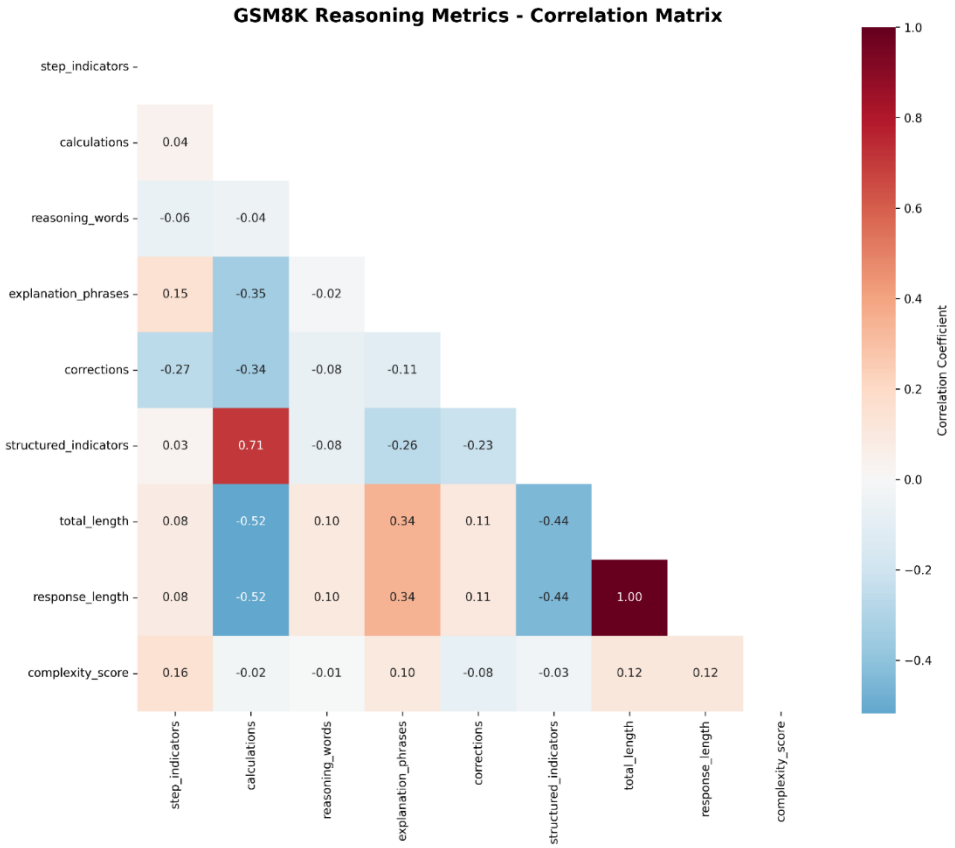

Figure 10: Reasoning metrics correlation matrix on GSM8K. Figure created by student researcher using Python in 2025.

We focus on the metrics of calculations and explanation phrases as our main measurements of calculative versus verbose reasoning. Calculations quantify concise, step-based computational work, whereas explanation phrases capture extended natural language exposition. These represent opposing reasoning modes: compact stepwise computation versus extended natural language elaboration. We quantify their impact on other selected metrics that measure structure, and output verbosity, as shown below in Table 3.

Table 3: Correlations between reasoning quality metrics. Table created by student researcher in 2025.

Metric 1	Metric 2	Pearson r	Pearson p-value	Spearman p-value
Calculations	Explanation Phrases	−0.350	3.53×10^{-30}	2.80×10^{-19}
Calculations	Structure Indicators	0.706	7.18×10^{-152}	1.84×10^{-93}
Calculations	Corrections	−0.341	1.09×10^{-28}	5.66×10^{-29}
Calculations	Response Length	−0.517	1.38×10^{-69}	1.57×10^{-83}
Explanation Phrases	Structured Indicators	−0.264	2.22×10^{-17}	8.51×10^{-9}
Explanation Phrases	Response Length	0.342	7.99×10^{-29}	1.09×10^{-26}

Our statistical analysis finds that calculations and explanation phrases are significantly anticorrelated with each other. Calculations positively align with structure indicators and negatively correlate with corrections, indicating that a calculations-heavy approach to reasoning results in a more logical chain of thought that is relatively error-free. Explanation phrases, however, negatively correlate with structure indicators and positively align with response length, suggesting that the explanation-heavy reasoning approach is more verbose and possibly less organized.

Importantly, by testing calculations against explanation phrases, we prove that reasoning is either calculations-heavy or explanations-heavy, and not both. In other words, we isolate two distinct reasoning modalities. Calculation-heavy traces tend to be short, structured, and error-free, whereas verbose traces are longer and dominated by explanatory scaffolding.

3.5.2 Mechanistic Identification

Our modality hypothesis holds true when tested mechanistically, and we identify certain features that track directly with one reasoning style over the other. We identify F59098 on the GSM8K dataset, and run a Pearson's correlation test as shown below in **Figure 11**. We find a statistically significant positive correlation between F59098 activations and calculations, and a statistically significant negative correlation between activation and explanation phrases. In other

words, we find that F59098 is a calculative-reasoning feature that prefers the computation modality and discourages verbose explanation.

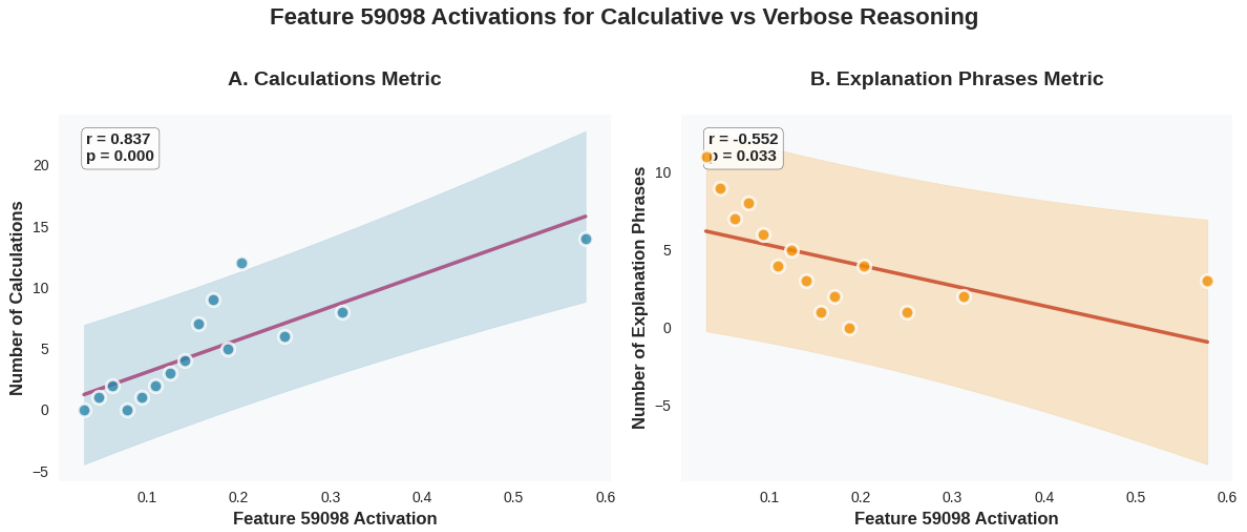

Figure 11: Correlation of Feature 59098 GSM8K activations with calculation and explanation phrase metrics. Figure created by student researcher using Python in 2025.

Interestingly, in harder settings new domain-specific reasoning modalities emerge. Observing the same Feature 59098 in the advanced Olympiads data, we observe that it maintains a slight pattern, but is not statistically significant as shown in Figure 12 below. The discrepancy between the grade-school and competition-math dataset indicates that F59908 is not conducive to advanced reasoning and only basic reasoning. We therefore isolate F59098 to a feature controlling calculative reasoning on lower difficulty problems.

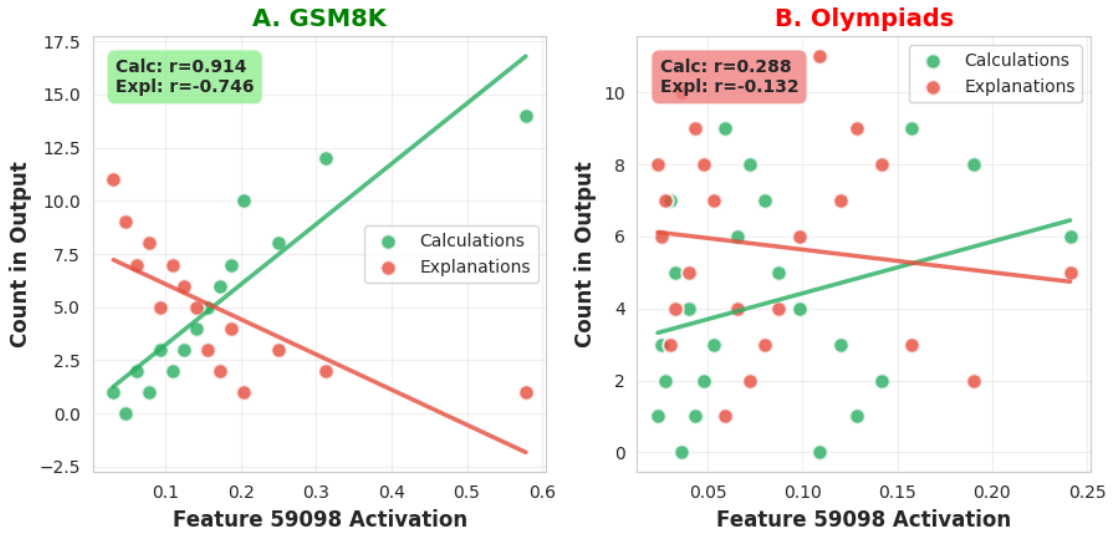

Figure 12: Correlation of Feature 59098 activation with calculative and explanatory reasoning on both GSM8K and Olympiad datasets. Figure created by student researcher using Python in 2025.

However, we still prove the existence of reasoning modality control features in the Olympiad dataset. Similar to how the dataset difficulty increase saw feature fracturing into specialized domains, we identify features that fracture into domain-specific modality controllers. We identify F35875, whose domain activations are shown below in Figure 13.

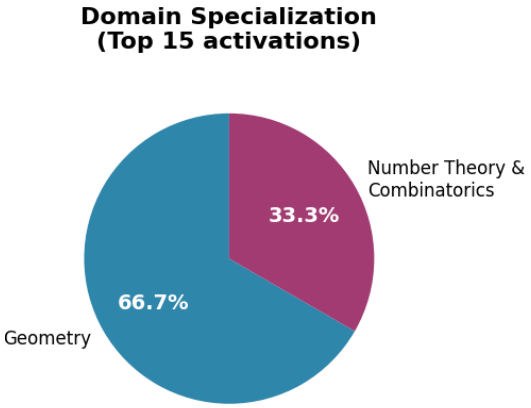

Figure 13: Domain distribution of top activations for Feature 35875. Figure created by student researcher using Python in 2025.

By analyzing Feature 35875's activations over domains, we identify it as an advanced geometry specialist. Through auto-interpretation as described in Section 2.3, we further classify F35875's specialization as "spatial geometry" with high activation on pyramid reasoning problems. Applying a similar statistical test for correlation as with F59098 in the geometry domain, we find that F35875 displays a statistically significant preference for the calculation-heavy reasoning modality over the verbose explanation modality within geometry, as shown below in Figure 14.

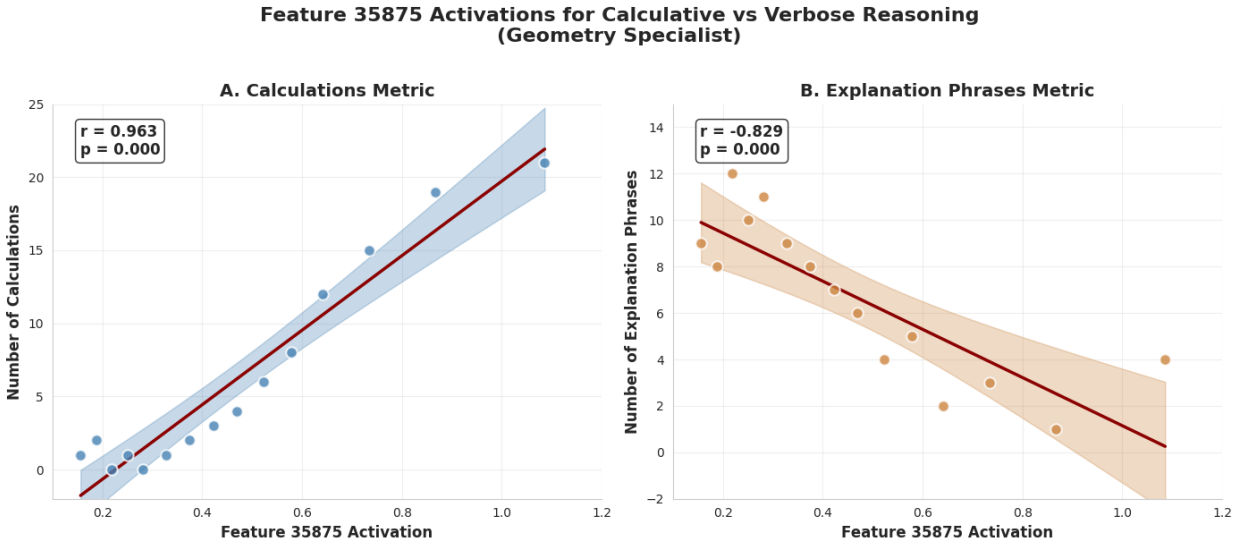

Figure 14: Correlation of Feature 35875 Olympiad geometry activations with calculation and explanation phrase metrics. Figure created by student researcher using Python in 2025.

The mechanistic identification of reasoning features and their statistical testing confirms our results from the correlation matrix from Figure 10. We find significant and clear evidence that LLM reasoning does not operate along a single dimension but instead splits into distinct modalities: a concise, calculation-driven style and a verbose, explanation-heavy style. Through correlation analysis of reasoning-quality metrics, we showed that these modalities are statistically opposed, with calculation-heavy traces being shorter and more structured, while explanation-heavy traces favor length and elaboration at the cost of organization. Crucially, our mechanistic analysis confirmed that specific features act as both generic and domain-specific modality controllers, selectively aligning with one reasoning style over the other. These findings provide the first mechanistic evidence that reasoning modalities are encoded within internal model features, proving that modality is not just an artifact of surface-level output but a structured and differentiable property of the model's internal representations.

4 Conclusion

4.1 Summary of Findings

This study develops a framework for mechanistic reasoning assurance by linking the internal representations of a large language model to quantifiable properties of reasoning quality. By combining sparse autoencoder (SAE) analysis with custom reasoning-quality metrics, we connect LLM internal feature activation structure with external behavioral evidence, providing a concrete basis for verifying whether reasoning occurs authentically inside the model rather than only in its text output.

We introduce task difficulty as a novel mechanistic variable. Using matched reasoning datasets of contrasting complexity (GSM8K and Olympiad Mathematics) under an identical model and SAE configuration, we show that features which appear generic on simpler problems fracture into domain-specific sub-features as cognitive load increases. This establishes difficulty as a controlled axis for probing the stability and specialization of reasoning features.

We provide the first mechanistic evidence of reasoning modalities within an LLM. We identify respectively a concise, calculation-driven modality and a verbose explanation-oriented modality. Each is governed by identifiable features whose activations predict the corresponding style of reasoning, demonstrating that reasoning strategy itself is encoded within latent space.

The integration of mechanistic interpretability with reasoning evaluation is a large step forward in quantitative LLM reasoning assurance. Correlations between specific interpretable

features and reasoning-quality metrics demonstrate that internal activations carry information about inference. Causal intervention experiments demonstrate that individual features actively control specific reasoning behaviors, establishing that reasoning is mechanistically governed by interpretable internal features. Holistically, we establish through both correlational and causal evidence that reasoning behavior in LLMs is mechanistically structured, difficulty-dependent, internally modular, and controllable through feature-level interventions. We establish a reproducible foundation for assessing not only whether a model reaches correct conclusions, but whether it reasons in a verifiable and intelligible way.

4.2 Implications

Large language models increasingly participate in domains that depend on sound reasoning: education, research, policy, and scientific communication. Despite this, their internal validity remains opaque. This study demonstrates that reasoning reliability can be examined mechanistically rather than inferred behaviorally. Mechanistic reasoning assurance transforms reasoning trustworthiness from an observed property into a verifiable one. This shift enables auditing of reasoning itself, not merely of outcomes.

For emerging AI governance frameworks, mechanistic verification of reasoning addresses a critical gap in current regulatory proposals. Standards requiring "explainable AI" typically focus on post-hoc output interpretations, which can be fabricated or misleading. Internal feature monitoring provides ground-truth verification: we can check whether claimed reasoning actually occurred inside the model. Our experiments demonstrate that these features are functional components that actively control reasoning outputs, enabling not just monitoring but potential intervention and correction of reasoning failures in deployed systems. This capability is essential for contexts where AI systems must demonstrate not just correct conclusions but legitimate inference processes, such as credit decisions, hiring algorithms, or medical recommendations.

The broader implication is that reasoning itself becomes an engineerable and auditable property of AI systems. Rather than treating reasoning as an emergent capability we can only measure indirectly, mechanistic interpretability enables direct observation and validation. As reasoning models become more capable and more widely deployed, the ability to verify reasoning integrity mechanistically will be essential for maintaining justified trust in AI systems that increasingly shape consequential human decisions.

4.3 Limitations and Future Work

This study examines a single model under controlled conditions. All findings are derived from DeepSeek-R1 Distill Llama-8B at a specific layer (block 19) using one SAE configuration. While this design ensures internal consistency, it limits direct generalizability to other LLMs. Although DeepSeek is representative of other models, we cannot conclude that these patterns exist in GPT-4, Claude, or Gemini. However, the mechanistic approach itself is architecture-independent and transfers to other transformer-based models.

Our findings are validated within arithmetic, algebra, geometry, and competition mathematics. While the feature specialization and modality distinction patterns we observe are likely general properties of how LLM features organize complex reasoning under difficulty, empirical validation in other reasoning domains is necessary. Causal reasoning, moral judgment, and probabilistic inference may engage different cognitive structures, and whether similar domain specialization emerges in those contexts remains an open empirical question. However, our central finding that reasoning features specialize under increased complexity is consistent with prior work on both feature interpretation and human cognitive specialization.

Future work should focus on testing the methodology introduced in this study across various LLM families. Applying this methodology to Claude, GPT-4o, Gemini, and open-source reasoning models such as the recently released GPT-OSS would test whether domain specialization and modality splitting are universal properties of reasoning-optimized LLMs or artifacts specific to DeepSeek's architecture and training. We predict these patterns will replicate across models given shared transformer foundations and similar reasoning optimization objectives, but further work is necessary.

Domain generalization to non-mathematical reasoning would test whether our framework is truly generic. The reasoning-quality metrics introduced here can be adapted to capture structure in causal chains, and analogical thinking. If similar specialization patterns emerge, such as features distinguishing deontological from utilitarian moral reasoning, this would support the hypothesis that complexity-driven specialization is a general principle of LLM cognition rather than a mathematical artifact.

If cross-model and cross-domain validation succeeds, this framework could mature into a standardized benchmark for reasoning assurance, enabling evaluation of AI systems not only by their conclusions but by the verifiable integrity of their internal reasoning processes.

References

1. Gokul, A. (2023). LLMs and AI: Understanding its reach and impact.
2. M. A. K. Raiaan et al., "A Review on Large Language Models: Architectures, Applications, Taxonomies, Open Issues and Challenges," in IEEE Access, vol. 12, pp. 26839-26874, 2024, doi: 10.1109/ACCESS.2024.3365742.
3. Ke, Z., Jiao, F., Ming, Y., Nguyen, X. P., Xu, A., Long, D. X., ... & Joty, S. (2025). A survey of frontiers in llm reasoning: Inference scaling, learning to reason, and agentic systems. *arXiv preprint arXiv:2504.09037*.
4. Shojaee, P., Mirzadeh, I., Alizadeh, K., Horton, M., Bengio, S., & Farajtabar, M. (2025). The illusion of thinking: Understanding the strengths and limitations of reasoning models via the lens of problem complexity. *arXiv preprint arXiv:2506.06941*.
5. Turpin, M., Michael, J., Perez, E., & Bowman, S. (2023). Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, *36*, 74952-74965.
6. Emmons, S., Jenner, E., Elson, D. K., Saurous, R. A., Rajamanoharan, S., Chen, H., ... & Shah, R. (2025). When chain of thought is necessary, language models struggle to evade monitors. *arXiv preprint arXiv:2507.05246*.
7. Lanham, T., Chen, A., Radhakrishnan, A., Steiner, B., Denison, C., Hernandez, D., ... & Perez, E. (2023). Measuring faithfulness in chain-of-thought reasoning. arXiv preprint arXiv:2307.13702.
8. Yu, T., Jing, Y., Zhang, X., Jiang, W., Wu, W., Wang, Y., ... & Tao, D. (2025). Benchmarking reasoning robustness in large language models. *arXiv preprint arXiv:2503.04550*.
9. Fu, J., Zhao, X., Yao, C., Wang, H., Han, Q., & Xiao, Y. (2025). Reward shaping to mitigate reward hacking in rlhf. *arXiv preprint arXiv:2502.18770*.
10. Korbak, T., Balesni, M., Barnes, E., Bengio, Y., Benton, J., Bloom, J., ... & Mikulik, V. (2025). Chain of thought monitorability: A new and fragile opportunity for ai safety. *arXiv preprint arXiv:2507.11473*.
11. Yao, Z., Liu, Y., Chen, Y., Chen, J., Fang, J., Hou, L., ... & Chua, T. S. (2025). Are Reasoning Models More Prone to Hallucination?. *arXiv preprint arXiv:2505.23646*.
12. Fodor, J. (2025). Line goes up? inherent limitations of benchmarks for evaluating large language models. *arXiv preprint arXiv:2502.14318*.

13. Chu, Z., Wang, S., Xie, J., Zhu, T., Yan, Y., Ye, J., ... & Wen, Q. (2025). Llm agents for education: Advances and applications. *arXiv preprint arXiv:2503.11733*.
14. Coz, P. L., Liu, J. A., Bhattacharjya, D., Curto, G., & Stinckwich, S. (2025). What Would an LLM Do? Evaluating Policymaking Capabilities of Large Language Models. *arXiv preprint arXiv:2509.03827*.
15. You, D., & Chon, D. (2024). Trust & Safety of LLMs and LLMs in Trust & Safety. *arXiv preprint arXiv:2412.02113*.
16. Burns, C., Ye, H., Klein, D., & Steinhardt, J. (2022). Discovering latent knowledge in language models without supervision. *arXiv preprint arXiv:2212.03827*.
17. Bereska, L., & Gavves, E. (2024). Mechanistic interpretability for AI safety--a review. *arXiv preprint arXiv:2404.14082*.
18. Cunningham, H., Ewart, A., Riggs, L., Huben, R., & Sharkey, L. (2023). Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600*.
19. O'Neill, C., Ye, C., Iyer, K., & Wu, J. F. (2024). Disentangling dense embeddings with sparse autoencoders. *arXiv preprint arXiv:2408.00657*.
20. Karvonen, A., Rager, C., Lin, J., Tigges, C., Bloom, J., Chanin, D., ... & Nanda, N. (2025). Saebench: A comprehensive benchmark for sparse autoencoders in language model interpretability. *arXiv preprint arXiv:2503.09532*.
21. Shu, D., Wu, X., Zhao, H., Rai, D., Yao, Z., Liu, N., & Du, M. (2025). A survey on sparse autoencoders: Interpreting the internal mechanisms of large language models. *arXiv preprint arXiv:2503.05613*.
22. Sharkey, L., Chughtai, B., Batson, J., Lindsey, J., Wu, J., Bushnaq, L., ... & McGrath, T. (2025). Open problems in mechanistic interpretability. *arXiv preprint arXiv:2501.16496*.
23. Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, Ł., Plappert, M., Tworek, J., Hilton, J., Nakano, R., Hesse, C., & Schulman, J. (2021). Training verifiers to solve math word problems (arXiv:2110.14168). arXiv. https://arxiv.org/abs/2110.14168.
24. Aslawliet. (n.d.). *Olympiads* Dataset. Hugging Face.
25. Chanin, D., Wilken-Smith, J., Dulka, T., Bhatnagar, H., Golechha, S., & Bloom, J. (2024). A is for absorption: Studying feature splitting and absorption in sparse autoencoders. *arXiv preprint arXiv:2409.14507*.

26. Elhage, N., Hume, T., Olsson, C., Schiefer, N., Henighan, T., Kravec, S., ... & Olah, C. (2022). Toy models of superposition. arXiv preprint arXiv:2209.10652.
27. Wagner, M. (2012). Number theory and the queen of mathematics. *The Mathematics Enthusiast*, *9*(1), 193-206.
28. Strachan, J.W.A., Albergo, D., Borghini, G. et al. Testing theory of mind in large language models and humans. Nat Hum Behav 8, 1285–1295 (2024). https://doi.org/10.1038/s41562-024-01882-z.
29. Lupita Estefania Gazzo Castañeda, Benjamin Sklarek, Dennis E. Dal Mas, Markus Knauff, Probabilistic and deductive reasoning in the human brain, NeuroImage, Volume 275, 2023, 120180, ISSN 1053-8119, https://doi.org/10.1016/j.neuroimage.2023.120180.