I Have Covered All the Bases Here: Interpreting Reasoning Features in Large Language Models via Sparse Autoencoders

Andrey V. Galichin[1,2,3], **Alexey Dontsov**[1,4],
Polina Druzhinina[1,3], **Anton Razzhigaev**[1,3],
Oleg Rogov[1,2,3], **Elena Tutubalina**[1,5,6], **Ivan Oseledets**[1,3]
[1]AIRI, Moscow, Russia
[2]MTUCI, Moscow, Russia
[3]Skoltech, Moscow, Russia
[4]HSE, Moscow, Russia
[5]ISP RAS Research Center for Trusted Artificial Intelligence, Moscow, Russia
[6]Sber AI, Moscow, Russia
a.v.galichin@mtuci.ru

Abstract

Recent LLMs like DEEPSEEK-R1 have demonstrated state-of-the-art performance by integrating deep thinking and complex reasoning during generation. However, the internal mechanisms behind these reasoning processes remain unexplored. We observe *reasoning* LLMs consistently use vocabulary associated with human reasoning processes. We hypothesize these words correspond to specific reasoning moments within the models' internal mechanisms. To test this hypothesis, we employ Sparse Autoencoders (SAEs), a technique for sparse decomposition of neural network activations into human-interpretable features. We introduce *ReasonScore*, an automatic metric to identify active SAE features during these reasoning moments. We perform manual and automatic interpretation of the features detected by our metric, and find those with activation patterns matching uncertainty, exploratory thinking, and reflection. Through steering experiments, we demonstrate that amplifying these features increases performance on reasoning-intensive benchmarks ($+2.2\%$) while producing longer reasoning traces ($+20.5\%$). Using the model diffing technique, we provide evidence that these features are present only in models with reasoning capabilities. Our work provides the first step towards a mechanistic understanding of reasoning in LLMs.

Code — https://github.com/AIRI-Institute/SAE-Reasoning
Extended version — https://arxiv.org/abs/2503.18878

Problem: Convert the point $(0, 3)$ in rectangular coordinates to polar coordinates. Enter your answer in the form (r, θ), where $r > 0$ and $0 \leq \theta < 2\pi$.

Generation (no steering)
Okay, so I have this problem where I need to convert the rectangular coordinates (0, 3) to polar coordinates.
$\{...$ **1500** tokens $...\} \leftarrow$ **Reasoning trace**
So, putting it all together, (0, 3) in rectangular coordinates is $\boxed{(3, \pi/2)}$ in polar coordinates.

Generation (steering)
Okay, so I have this problem here where I need to convert the point (0, 3) from rectangular (which is the same as Cartesian) coordinates to polar coordinates.
$\{...$ **2000** tokens $...\} \leftarrow$ **Increased reasoning trace**
I think I've covered all the bases here. Calculated r, determined θ, checked using different methods, and even considered the quadrant placement. I don't see any issues with the reasoning. So, I feel confident that the polar coordinates for the point (0, 3) are $\boxed{(3, \pi/2)}$.

Figure 1: Illustration of steering (amplifying) reasoning-specific features during LLM generation. Default generation (*no steering*) shows standard model reasoning, whereas *steering* induces increased reasoning, self-correction, and graceful transition to the final answer—evidence that the identified features are responsible for the reasoning concept.

1 Introduction

Large Language Models (LLMs) have achieved remarkable success in natural language processing (Brown et al. 2020), evolving beyond simple token prediction tasks towards explicit reasoning behaviors, such as step-by-step problem-solving (Wei et al. 2022; Kojima et al. 2022; Wang et al. 2022) and self-reflection (Madaan et al. 2023; Shinn et al. 2023). Recently, specialized models which we denote as *reasoning* models, such as OpenAI's o1 (OpenAI 2024) and DEEPSEEK-R1 (Guo et al. 2025), have significantly improved performance on complex reasoning tasks. Trained through advanced fine-tuning and reinforcement learning (Shao et al. 2024), these models incorporate reasoning and reflective problem-solving by generating long chains of thought before providing final answers. These advances raise a new research question: How are such reasoning ca-

pabilities internally encoded within LLMs?

A growing body of work suggests that LLMs represent meaningful concepts as linear directions in their activation spaces (Mikolov et al. 2013; Elhage et al. 2022; Park, Choe, and Veitch 2023; Nanda, Lee, and Wattenberg 2023; Jiang et al. 2024). However, identifying these directions remains challenging. SAEs offer a principled approach to disentangle activations into sparse, interpretable *features* (Huben et al. 2024; Gao et al. 2024b; Templeton 2024; Marks et al. 2024). Given a trained SAE, the interpretation of its features could be performed by activation analysis (Bricken et al. 2023), targeted interventions (Templeton 2024), or automated methods (Paulo et al. 2024; Kuznetsov et al. 2025). While SAEs have proven effective in discovering features for various concepts (Shu et al. 2025), their ability to isolate reasoning-specific features remains unexplored.

In this work, we investigate whether reasoning processes in *reasoning* LLMs can be identified and decomposed into interpretable directions within their activation spaces. We analyze the outputs produced by these models, and find a consistent pattern in which they employ words associated with human reasoning processes: uncertainty (e.g. "perhaps"), reflection (e.g. "however"), and exploration (e.g. "alternatively") (Chinn and Anderson 1998; Boyd and Kong 2017; Gerns and Mortimore 2025). We hypothesize that these linguistic patterns correspond to the moments of reasoning within the models' internal mechanisms. To test this, we construct a vocabulary of reasoning words. We then use SAEs to decompose LLM activations into interpretable features and propose `ReasonScore`, a metric that quantifies the degree to which a given SAE feature is active on the reasoning vocabulary.

We evaluate the features identified by `ReasonScore` using manual (Bricken et al. 2023) and automatic interpretation (Kuznetsov et al. 2025) techniques, and find the set of 46 features that demonstrate interpretable activation patterns corresponding to uncertainty, exploratory thinking, and reflection. We perform steering experiments (Fig. 1) and show that amplifying these reasoning features leads to improved performance on reasoning-intensive benchmarks ($+13.4\%$ on AIME-2024, $+2.2\%$ on MATH-500, and $+4\%$ on GPQA Diamond) while producing longer reasoning traces ($+18.5\%$ on AIME-2024, $+20.5\%$ on MATH-500, and $+13.9\%$ on GPQA Diamond). Through model diffing (Bricken et al. 2024), we demonstrate that these reasoning features emerge only in *reasoning* LLMs and are absent in base models. Our results provide mechanistic evidence that specific, interpretable components in LLMs representations are causally linked to reasoning behavior.

The contributions of this paper are the following:

- We introduce `ReasonScore`, an automatic metric to identify the SAE features responsible for reasoning and confirm its effectiveness using interpretability techniques.
- We provide causal evidence from steering experiments, demonstrating that amplifying identified features induces reasoning behavior.
- We analyze the emergence of reasoning features in LLMs through model diffing technique, and confirm their existence only after the *reasoning* fine-tuning stage.

2 Interpretability with SAEs

SAEs aim to learn a sparse decomposition of model activations to identify disentangled features that correspond to meaningful concepts (Bricken et al. 2023). Here, a *feature* refers to an individual component of the learned representation that captures specific, human-interpretable characteristics of the input data.

The core idea behind SAEs is to reconstruct model activations $x \in \mathbb{R}^n$ as a sparse linear combination of learned feature directions, where the feature *dictionary* dimensionality $m \gg n$. Formally, we extract LLM activations from some intermediate state in the model and train a two-layer autoencoder:

$$\begin{aligned} f(x) &= \sigma(W_{\text{enc}}x + b_{\text{enc}}), \\ \hat{x}(f) &= W_{\text{dec}}f + b_{\text{dec}}. \end{aligned} \tag{1}$$

Here, $f(x) \in \mathbb{R}^m$ is a sparse vector of feature magnitudes and $\hat{x}(f) \in \mathbb{R}^n$ is a reconstruction of the original activation x. The columns of W_{dec}, which we denote by $W_{\text{dec},i}$, $i = 1, \ldots, m$, represent the dictionary of directions, or *features*, into which the SAE decomposes x. The activation function σ enforces non-negativity in $f(x)$.

The training objective used to train SAEs minimizes a reconstruction loss $\mathcal{L}_{\text{recon}}$ and an additional sparsity-promoting loss $\mathcal{L}_{\text{sparsity}}$. This objective forces SAE to learn a small set of interpretable features that capture the distinct properties of the activations.

In our work, we use vanilla SAE (Bricken et al. 2023) with `ReLU` activation function. Following (Conerly et al. 2024), we use a squared error reconstruction loss and a modified L1 penalty as a sparsity loss:

$$\mathcal{L} = \underbrace{\|x - \hat{x}\|_2^2}_{\mathcal{L}_{\text{recon}}} + \lambda \underbrace{\sum_{i=1}^{m} f_i \left\|W_{\text{dec},i}\right\|_2}_{\mathcal{L}_{\text{sparsity}}}, \tag{2}$$

where λ is the sparsity penalty coefficient.

3 Method

We identify reasoning-specific features through a two-step approach. First, we examine the language space of reasoning words used by *reasoning* LLMs, and construct the respective vocabulary $\mathcal{R}$ (Sec. 3.1). Secondly, we introduce `ReasonScore` to find the sparse autoencoder features responsible for reasoning capabilities (Sec. 3.2).

3.1 Reasoning Vocabulary

Reasoning words are linguistic features associated with exploratory talk as humans talk-to-learn, explore ideas, and probe each other's thinking (Boyd and Kong 2017).

In the original DEEPSEEK-R1 paper (Guo et al. 2025), the authors demonstrated that the model spontaneously exhibits sophisticated human-like behaviors, such as reflection, where it revisits and reevaluates its previous steps, and exploration of alternative problem-solving approaches. In particular, the model explicitly employs words that mirror

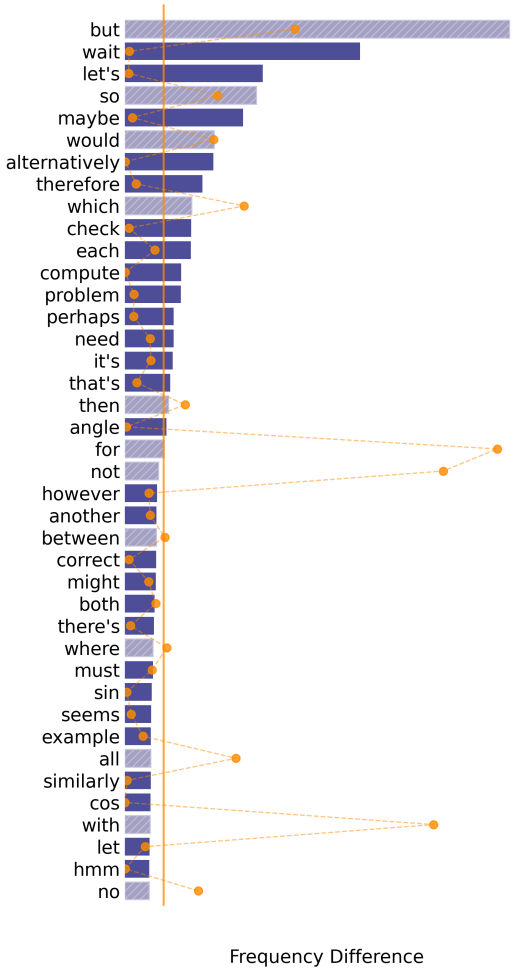


Figure 2: The distribution of top 40 words with the greatest change in frequency between reasoning traces of DeepSeek-R1 and ground-truth solutions of math problems. Orange dots show the frequency from Google Books Ngram Corpus. We remove the words with absolute frequency above the pre-defined threshold (orange line), and keep those with the high relative frequency indicating reasoning.

the introspective language humans use when thinking (such as "maybe", "but", "wait"). We hypothesize that these moments correspond directly to the internal reasoning process of the models, which is consistent with studies on human thinking (Chinn and Anderson 1998; Boyd and Kong 2017).

To extract the models' reasoning vocabulary, we use an approach similar to that of (Rayson and Garside 2000). We construct two corpora from the OPENTHOUGHTS-114K (Thoughts 2025) dataset: ground-truth samples containing formal and step-by-step solutions to the problems, and the solutions obtained using DEEPSEEK-R1 for the same problems. For each word, we calculate its frequency in the tasks solutions p_{solution} and in the thinking solutions p_{think}, then sort all words by the frequency difference $p_{\text{think}} - p_{\text{solution}}$. Next, we select the top-k words by frequency difference, where k is determined by the point where the frequency distribution plateaus, and filter out words with high presence in the Google Books Ngram Corpus (Michel et al. 2011) (Fig. 2). To determine the final vocabulary from this candidate set, we choose words that best capture reasoning behavior. This includes words that match those considered in the linguistic literature (Chinn and Anderson 1998; Boyd and Kong 2017) and those that we identify through manual analysis of model traces exhibiting reasoning patterns.

Following this pipeline, we select 10 words indicating reasoning as models' *reasoning* vocabulary and denote it by $\mathcal{R}$. The exact list of words can be found in Appx. A.1. Ablation experiments confirm these words play a functional role in reasoning capabilities (see Sec. 4.3 for setup, Appx. A.2 for results).

3.2 ReasonScore

To find SAE features that capture reasoning-related behavior, we follow our hypothesis and introduce `ReasonScore`, which measures the contribution of i-th feature to reasoning. Using a dataset of model's activations (see details in Sec. 4.1) $\mathcal{D} = \mathcal{D}_{\mathcal{R}} \cup \mathcal{D}_{\neg\mathcal{R}}$, where $\mathcal{D}_{\mathcal{R}}$ contains token activations corresponding to words in $\mathcal{R}$ and $\mathcal{D}_{\neg\mathcal{R}}$ contains all other activations, we first define a score:

$$s_i = \frac{\mu(i, \mathcal{D}_{\mathcal{R}})}{\sum_j \mu(j, \mathcal{D}_{\mathcal{R}})} - \frac{\mu(i, \mathcal{D}_{\neg\mathcal{R}})}{\sum_j \mu(j, \mathcal{D}_{\neg\mathcal{R}})}, \tag{3}$$

where $\mu(i, \mathcal{D}) = \frac{1}{|\mathcal{D}|} \sum_{x \in \mathcal{D}} f_i(x)$ is the average activation value of the i-th feature on dataset $\mathcal{D}$. This score is similar to the one in (Cywiński and Deja 2025) and identifies features that concentrate the most of their activation mass on reasoning words.

However, analysis of feature activations only on individual words may miss important contextual information. The words in $\mathcal{R}$ are critical indicators of the reasoning process and also serve as transition points, signaling shifts in the thought process, uncertainty, or reflection. Therefore, a feature involved in reasoning should activate not only on the reasoning words, but also as the model approaches and continues through these transitions. To capture it, we define $\mathcal{D}^{\text{W}}_{\mathcal{R}}$ as the dataset that contains activations within a fixed-width context window around tokens corresponding to words in $\mathcal{R}$, and $\mathcal{D}^{\text{W}}_{\neg\mathcal{R}}$ contains all other activations. We modify Eq. 3 to use the new version of the datasets.

To penalize features that activate only on a small fraction of $\mathcal{R}$, we further introduce an *entropy penalty*. For i-th feature, we first calculate $\mu(i, \mathcal{D}^{\text{W}}_{r_j})$ for each word $r_j \in \mathcal{R}$, normalize these values into a probability distribution $p_i(r_j) = \frac{\mu(i, \mathcal{D}^{\text{W}}_{r_j})}{\sum_{k \in \mathcal{R}} \mu(i, \mathcal{D}^{\text{W}}_{r_k})}$, and compute the entropy:

$$\text{H}_i = -\frac{1}{\log |\mathcal{R}|} \cdot \sum_{j}^{|\mathcal{R}|} p_i(r_j) \log p_i(r_j). \tag{4}$$

Here, $\log |\mathcal{R}|$ normalizes the entropy to $[0, 1]$, with $\text{H}_i = 1$ indicating perfect uniformity over $\mathcal{R}$. By adding the entropy

Feature #4395

But how to verify this.\n\nLet me re-express the inequalities hold. But wait, maybe the constant up to some equivalence.\n\nBut let me check if there So this is unclear.\n\nBut looking at the sample

Feature #25953

rahedra.\n\nWait, here's another idea. (71 pm). Wait, maybe different sources have slightly formula has a different sign.\n\nAlternatively, maybe the coefficient low. Let me verify. Another way: I remember

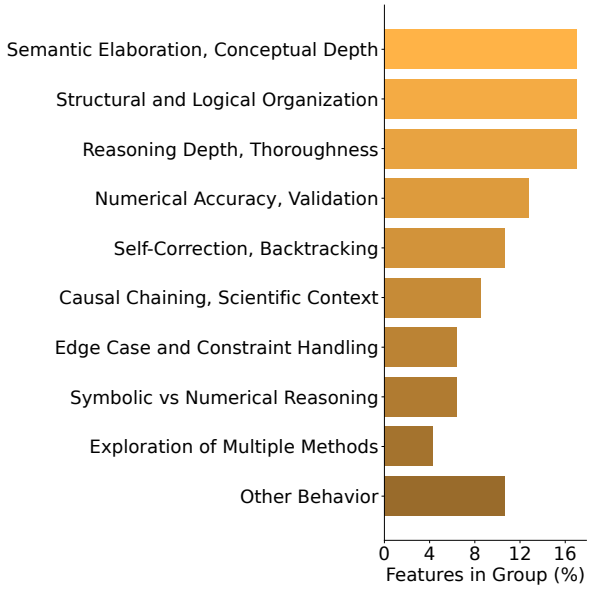
Feature #46691

mass ranges. Hmm.\n\nNow, detection. LIGO the starting points. That seems a bit counterintuitive trigonometric identities. Hmm, but maybe a synthetic regardless of starting values. Interesting.\n\nLet me try

Feature #61104

some code?), but that's not standard. Alternatively is in a different row. That doesn't make sense That's a stretch, but perhaps that's the connection , then maybe more. Wait, this is getting complicated

(a) Top-activating examples from the manually verified set of features.



(b) Distribution of manually verified set of features on function groups generated by GPT-4o.

Figure 3: Interpretability results for manually verified set of features in our SAE: (a) Examples of feature interfaces used in manual interpretation experiments, (b) Distribution of reasoning features on function groups obtained by automatic interpretation pipeline by using GPT-4o as a judge.

penalty in Eq. (3), we define the $\texttt{ReasonScore}$ for the i-th SAE feature as:

$$\texttt{ReasonScore}_i = \frac{\mu(i, \mathcal{D}^{\text{W}}_{\mathcal{R}})}{\sum_j \mu(j, \mathcal{D}^{\text{W}}_{\mathcal{R}})} \cdot \text{H}_i^{\alpha} - \frac{\mu(i, \mathcal{D}^{\text{W}}_{\neg\mathcal{R}})}{\sum_j \mu(j, \mathcal{D}^{\text{W}}_{\neg\mathcal{R}})}. \quad (5)$$

where α controls the trade-off between specificity ($\alpha \to 0$) and generalization ($\alpha > 1$).

We identify the set of reasoning features in a SAE based on their $\texttt{ReasonScore}$ and define the corresponding set of feature indices as:

$$\mathcal{F}_{\mathcal{R}} = \{i \mid i \in [1, m],\ \texttt{ReasonScore}_i > \tau\}, \quad (6)$$

where τ is the q-th quantile of the $\texttt{ReasonScore}$ distribution across all features.

4 Evaluation

In this section, we analyze how effectively our discovered features model reflection, uncertainty, and exploration within the *reasoning* model. We discuss our experimental setup (Sec. 4.1), perform manual and automatic interpretation of the features we find (Sec. 4.2), and conduct steering experiments with these features on various benchmarks (Sec. 4.3). Finally, we apply the model diffing technique to demonstrate that these features exist only in models with reasoning capabilities (Sec. 4.4).

4.1 Experimental Setup

Model. We apply SAE to the output activations from the 19-th layer of the DEEPSEEK-R1-LLAMA-8B model. This model was selected for its *reasoning* capabilities and open-source availability. The 19-th layer ($\approx 60\%$ model depth) was chosen because at this point LLMs predominately store the most of their knowledge (Chen et al. 2023; Jin et al. 2025). We provide results for other layers of DEEPSEEK-R1-LLAMA-8B and another model family in Appx. D.

Data. We train SAE on the activations of the model generated using text data from the LMSYS-CHAT-1M (Zheng et al. 2023) and OPENTHOUGHTS-114K (Thoughts 2025) datasets. The first provides a broad and diverse spectrum of real-world text data, which we denote as *base data*, while the latter provides high-quality reasoning traces generated by DEEPSEEK-R1 for math, science, code, and puzzle samples, which we denote as *reasoning data*. The SAE is trained on 1B tokens, evenly split between the two datasets, with a context window of 1,024 tokens.

Training. We set the SAE dictionary dimensionality to $m = 65{,}536$, which is 16 times larger than the model activation size $n = 4{,}096$ following established practices (Lieberum et al. 2024), and adopt the same training settings as in the Anthropic April update (Conerly et al. 2024). We train with the Adam optimizer (Kingma and Ba 2014) with $(\beta_1, \beta_2) = (0.9, 0.999)$, batch size of 4,096, and a learning rate $\eta = 5 \times 10^{-5}$. The learning rate is decayed linearly to zero over the last 20% of training. The gradient norm is

clipped to 1. We use a linear warmup for the sparsity coefficient from $\lambda = 0$ to $\lambda = 5$ over the first 5% training steps.

Evaluation. We use the mean L0-norm of latent activations, $\mathbb{E}_x \|f(x)\|_0$, as a measure of sparsity. To measure reconstruction quality, we use fraction of variance of the input explained by the reconstruction. Both metrics were computed on $2{,}048$ sequences of length $1{,}024$.

At a L0 of 85 the reconstruction of our SAE explains 84.2% of the variance in model activations. This shows that our SAE achieves reliable reconstruction performance at a low sparsity level, allowing a decomposition of raw activations into interpretable features.

`ReasonScore`. We calculate `ReasonScore` (Eq. 5) on 10M tokens from the OPENTHOUGHTS-114K dataset. To collect $\mathcal{D}^{\text{W}}_{\mathcal{R}}$, we use an asymmetric window with 2 preceding and 3 subsequent tokens, following established practices in keyphrase extraction (Mihalcea and Tarau 2004; Breidt 1993; Zhang et al. 2020). We set $\alpha = 0.7$ for the *entropy penalty* as a reasonable default. Based on the empirical analysis of `ReasonScore` distribution (see Appx. A.3), we set $q = 0.997$ in Eq. (6), resulting in $|\mathcal{F}_{\mathcal{R}}| = 200$ features.

4.2 Interpretability of Reasoning Features

Manual Interpretation. To evaluate the features identified by `ReasonScore`, we manually interpret each of the 200 features in $\mathcal{F}_{\mathcal{R}}$. For each feature, we find the examples in a subset of the OPENTHOUGHTS-114K corpus that caused the feature to activate, and construct the interface proposed in (Bricken et al. 2023). This mainly includes examples of when the feature activates, its effect on the logits when it does, and other statistics. We determine whether a feature qualifies as a good reasoning candidate if: (1) when it is active, the relevant concept is reliably present in the context, (2) it triggers in various examples of reasoning tasks, and (3) its activation impacts interpretable logits that correspond to reasoning processes.

Through our analysis, we identify three behavioral modes that characterize models' reasoning process:

- **Uncertainty**: Moments where the model exhibits hesitation, doubts, and provisional thinking
- **Exploration**: Moments where the model considers multiple possibilities, connects ideas, examines different perspectives
- **Reflection**: Moments where the model revisits and reevaluates its previous steps

Our manual analysis reveals a set of 46 features that exhibit these patterns, which we believe are responsible for the reasoning mechanisms of the model. We denote this set by $\mathcal{F}^{\text{manual}}_{\mathcal{R}} \subset \mathcal{F}_{\mathcal{R}}$. In Fig. 3a, we provide examples of feature interfaces used for interpretation. The results demonstrate features that consistently activate in contexts representing model's uncertainty ($\#61104$), exploration ($\#25953$), and reflection ($\#4395, \#46691$). Additional examples of interfaces can be found in Appx. B.1.

Feature #	AIME 2024		MATH-500		GPQA Diamond	
	maj@4	K	maj@4	K	maj@4	K
NS	53.3	12.4	93.2	3.9	50.0	7.9
3942	56.7	11.1	93.0	3.4	46.5	6.7
4395	56.7	**14.7**	**95.4**	4.1	52.0	8.5
16441	60.0	14.0	95.0	4.1	54.0	8.3
16778	56.7	14.1	94.0	**4.7**	51.0	**9.0**
25953	60.0	12.8	94.2	4.2	53.0	8.1
46691	56.7	14.0	94.2	4.2	**54.0**	8.0
61104	**66.7**	12.0	95.0	3.6	53.0	7.5

Table 1: Performance and average number of output tokens (K) for different steering experiments on reasoning-related benchmarks. NS stands for "No steering".

Automatic Interpretation. To complement our manual analysis, we annotate these features with an automatic interpretation pipeline (Kuznetsov et al. 2025). This approach employs feature steering, a technique that modulates feature activations to analyze their functional influence. For each i-th feature, we estimate its maximum activation $f_i^{\max}$ using a subset of the OPENTHOUGHTS-114K corpus. During response generation, we modify model activations as follows:

$$x' = x + \gamma f_i^{\max} W_{\text{dec},i}, \tag{7}$$

where γ controls the steering strength.

To evaluate the impact of i-th feature on reasoning capabilities, we generate multiple outputs by varying $\gamma \in [-4, 4]$, pass them to GPT-4O, and ask it to generate an explanation or function that best describes the semantic influence caused by steering a feature. The result, shown in Fig. 3b, reveals that the features we found group into distinct reasoning-related patterns. Only a small fraction of features from $\mathcal{F}^{\text{manual}}_{\mathcal{R}}$ (5) was assigned to a mixed class "Other Behavior" containing mixed explanation. We provide a more comprehensive description of auto-interpretability pipeline results in Appx. B.2.

Takeaway 1. Manual interpretation experiments confirm that `ReasonScore` identifies features that describe model's reasoning capabilities, revealing 46 features that represent uncertainty, exploration, and reflection. Automatic interpretation demonstrates that these features are causally linked to reasoning behavior.

4.3 Steering Reasoning Features

To demonstrate whether our interpretations of features describe their influence on model behavior, we further experiment with feature steering.

Our goal is to verify if steering reasoning features improve the LLM's performance on reasoning-related benchmarks. Following the setup in DEEPSEEK-R1, we evaluate performance on AIME 2024 (MAA 2024), MATH-500 (Hendrycks et al. 2021), and GPQA Diamond (Rein et al. 2023). To obtain steering results for i-th feature, we modify the activations during response generation according to

Eq. (7). To determine the optimal steering strength that can influence model outputs without significantly damaging capabilities, we ran evaluations with a small subset of 10 reasoning features on MATH-500. We varied the steering strength γ from 1 to 8. Based on these experiments, we determined the optimal range $\gamma \in [1, 3]$, which aligns with the findings in (Durmus et al. 2024). For all subsequent experiments, we set the steering strength $\gamma = 2$.

We perform a preliminary analysis to identify the most promising features for reasoning enhancement from our set of manually chosen features $\mathcal{F}_{\mathcal{R}}^{\text{manual}}$. For each feature, we measure the accuracy (or pass@1 (Chen et al. 2021)) on MATH-500 and evaluate the results. Of the 46 features, 9 improve performance by $\geq 0.5\%$, 29 show no or minimal performance degradation (≤ 2.0), and the remaining 8 decrease performance by at most 4%. Interestingly, we identify feature #3942, which produces substantially shorter responses while maintaining negligible performance degradation. For further analysis, we select the 9 top-performing features and feature #3942.

We evaluate these 10 features across all reasoning benchmarks. We report majority voting (Snell et al. 2024; Brown et al. 2024) across 4 responses and the average number of tokens generated during the model's thinking process. The results, shown in Tab. 1, demonstrate that steering 7 out of 10 features produces consistent improvements in both performance and reasoning depth. Feature #61104 yields the most significant performance gain on AIME-2024 (+13.4%). Feature #16778 produces the longest reasoning traces on average (+13.7% on AIME-2024, +20.5% on MATH-500, and +13.9% on GPQA Diamond) and consistently outperforms the "no steering" baseline. Feature #3942 produces shortest reasoning traces on average (−7.7%) with minor performance degradation. We provide examples of generated solutions without and with feature steering in Appx. C.

Takeaway 2. We find that amplifying certain reasoning features prolongs the internal thought process and correlates with increased performance on reasoning-related tasks.

4.4 Stage-wise Emergence of Reasoning Features

Our interpretation experiments (Sec. 4.2) revealed that features identified by `ReasonScore` exhibit activation patterns consistent with reasoning processes. The steering experiments (Sec. 4.3) provided causal evidence by demonstrating that amplification of these features improves performance on reasoning-intensive benchmarks. Given these findings, we now aim to answer the next important question: do these reasoning features naturally emerge during standard pre-training procedure, or are they specifically induced by the *reasoning* fine-tuning process?

To answer this question, we use the stage-wise fine-tuning (FT) technique proposed in (Bricken et al. 2024). This approach aims to isolate how features evolve across different model and dataset combinations. In our experiments, we examine how the features change between two model states: before (*base model*) and after (*reasoning model*) *reasoning* fine-tuning stage. We accomplish this by training a SAE on the base model before it has been fine-tuned, and then fine-tuning it on either the reasoning model or the fine-tuning data. Formally, we define four distinct stages:

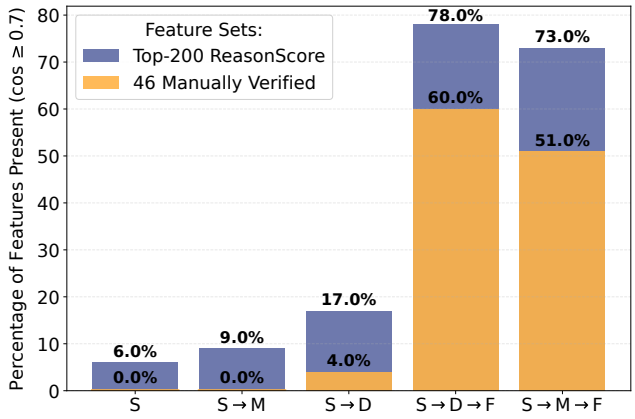


Figure 4: Percentage of reasoning features present at each stage of the diffing pipeline. The blue bars represent the features from $\mathcal{F}_{\mathcal{R}}$, the orange bars represent the $\mathcal{F}_{\mathcal{R}}^{\text{manual}}$ features. Feature is considered present if its cosine similarity with any feature in corresponding stage's SAE is ≥ 0.7. Stages: (S) Base model + base data; (S→D) Base model + reasoning data; S→M Reasoning model + base data; (S→D/M→F) Reasoning model + reasoning data.

Stage S: base model + base data (starting point).

Stage D: base model + reasoning data (isolating dataset effects).

Stage M: reasoning model + base data (isolating model effects).

Stage F: reasoning model + reasoning data (full FT).

We analyze these changes through two fine-tuning trajectories, each involving two sequential fine-tuning stages: (1) **S→D→F** takes initial SAE (Stage **S**), fine-tunes it on reasoning data (**S→D**), and finally fine-tunes on both reasoning model and reasoning data (**D→F**); (2) **S→M→F** takes initial SAE (Stage **S**), fine-tunes it on reasoning model (**S→M**), and finally fine-tunes on both reasoning model and reasoning data (**M→F**). If reasoning features are present only in *reasoning* models, we should observe the emergence of these features in response to **both** reasoning model and reasoning data (Stage **F**). This corresponds to the final steps of the fine-tuning trajectories: (**S→D/M→F**).

We use LLAMA-3.1-8B (Grattafiori et al. 2024) as the base model and SLIMPAJAMA (Soboleva et al. 2023) as base data, chosen over LMSYS-CHAT-1M because it better matches the model's pre-training distribution of LLAMA-3.1-8B. Each stage follows the setup in Sec. 4.1, with fine-tuning stages consuming 30% of the tokens needed for training from scratch. For each i-th feature in $\mathcal{F}_{\mathcal{R}}$, we assess its presence at each stage by cosine similarity (cos) between feature vectors, considering it present if $\cos \geq 0.7$ with any feature in the stage's SAE (Bricken et al. 2024).

Fig. 4 shows the percentage of reasoning features present at each fine-tuning stage. We find that the reasoning fea-

tures are almost absent in the base model and after switching to the reasoning model (0% of manually verified features $\mathcal{F}_{\mathcal{R}}^{\text{manual}}$). When introducing the reasoning data to the base model (**S→D**), only 4% of the verified reasoning features emerge, indicating that exposure to the reasoning content alone is insufficient to develop these features. Finally, when we incorporate both the reasoning data and the reasoning model, we observe that 60% of the verified reasoning features appear in the (**S→D→F**) stage and 51% in the (**S→M→F**) stage. The noticeable increase in the presence of features only when both reasoning data and model are combined provides compelling evidence that `ReasonScore` identifies features associated with the model's reasoning processes rather than general capabilities.

Takeaway 3. We show that most of the features found by `ReasonScore` emerge only after the *reasoning* fine-tuning stage. Exposure to the reasoning data or reasoning model alone is insufficient to develop these features.

5 Related Work

5.1 Mechanistic Interpretability (MI)

Various methods exist to shed light on the inner workings of LLMs, including attention analysis (Vaswani et al. 2017), gradient-based methods (Simonyan, Vedaldi, and Zisserman 2014), and probing techniques that offer insights into the information captured within different layers of an LLM (Alain and Bengio 2016). MI techniques such as activation patching (Meng et al. 2022) and feature steering (Cao et al. 2024; Soo, Teng, and Balaganesh 2025) aim to reverse-engineer and control model behavior. The logit lens provides a way to observe the model's token predictions at different processing stages (Nostalgebraist 2020).

5.2 Sparse Autoencoders

SAEs have emerged as a key tool for understanding the internal representations of LLMs (Gao et al. 2024a; Huben et al. 2024). By learning a sparse decomposition of model activations, SAEs identify disentangled features that correspond to meaningful concepts (Marks et al. 2024).

SAE features are significantly more monosemantic than individual neurons, making them effective for MI (Leask et al. 2025). A key challenge in using SAEs for MI is ensuring that the extracted features are monosemantic and robust. Yan et al. (Yan et al. 2024) propose using feature decorrelation losses to enforce better separation between learned latents, preventing redundancy. Recent advances in cross-layer SAEs (Shi et al. 2025) enable the analysis of more abstract, high-level reasoning patterns across multiple layers.

SAEs are valuable for analyzing model development across training. Crosscoders (Lindsey et al. 2024) map features across checkpoints, while stage-wise diffing (Bricken et al. 2024) compares SAEs trained at different stages. We use diffing for its efficiency and simplicity, extending prior sleeper-agent analyses to reasoning behavior.

5.3 Reasoning LLMs

Recent LLM innovations have focused on models with explicit reasoning abilities, including OpenAI's o1 (OpenAI 2024), DEEPSEEK-R1 (Guo et al. 2025), and QWQ-32B-PREVIEW (Team 2024). These methods employ rule-based reinforcement learning using correctness scores (final answer accuracy) and format scores (output structure compliance), leading to advanced reasoning behaviors like self-correction and reflection, denoted as an "aha moment" in the DEEPSEEK-AI report (Guo et al. 2025).

Despite the success of rule-based reinforcement learning in enabling reasoning capabilities, how these models encode their internal reasoning remains unclear. We address this problem using SAEs to find interpretable features responsible for underlying reasoning mechanisms, which, to the best of our knowledge, has not been done yet.

6 Conclusion

In this work, we present a novel methodology for uncovering reasoning mechanisms in LLMs through SAEs. We introduce `ReasonScore`, a metric that identifies reasoning-related SAE features based on their activation patterns using a curated introspective vocabulary. Manual and automatic MI reveal features corresponding to uncertainty, exploratory thinking, and self-reflection. Through steering experiments, we provide causal evidence that certain features selected by `ReasonScore` directly correspond to the model's reasoning behaviors. Amplifying them prolongs the internal thought process and increases performance on reasoning-related benchmarks. Stage-wise analysis confirms that these features emerge only after reasoning fine-tuning. Our work provides the first mechanistic evidence that specific, interpretable components of LLM representations are causally linked to complex reasoning behaviors.

7 Limitations

ReasonScore. Our metric depends on hyperparameters (window size, entropy penalty α) that require further ablation studies. Of the 200 candidates, we found 46 interpretable features. Although other features can also contribute to reasoning, we could not confidently classify them due to ambiguous activation patterns. Finally, our reasoning vocabulary may not capture all reasoning patterns. These limitations suggest opportunities for future work.

SAEs. SAEs provide a powerful interpretability framework. However, it suffers from problems that complicate the extraction of fully interpretable features (Chanin et al. 2024; Leask et al. 2025). This may cause us to miss some features.

Emergence of Reasoning Features. Although the results in Sec. 4.4 support our hypothesis, we acknowledge certain limitations of the diffing approach. The cosine similarity threshold (0.7) is empirically chosen following the initial work, and may miss similar features if the representation is rotated during one of the fine-tuning stages. Only 60% of the verified features (and 78% of the $\mathcal{F}_{\mathcal{R}}$ features) appeared in the final stage, likely due to fine-tuning SAE rather than training from scratch. These limitations show that our approach can result in false negative and false positive predictions. However, we believe that our primary finding remains valid even under these limitations.

Acknowledgments

This work was supported by a grant, provided by the Ministry of Economic Development of the Russian Federation (agreement identifier 000000C313925P4G0002) and the agreement with the Ivannikov Institute for System Programming of the Russian Academy of Sciences dated June 20, 2025 No. 139-15-2025-011.

References

Alain, G.; and Bengio, Y. 2016. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*.

Boyd, M.; and Kong, Y. 2017. Reasoning Words as Linguistic Features of Exploratory Talk: Classroom Use and What It Can Tell Us. *Discourse Processes*, 54(1): 62–81.

Breidt, E. 1993. Extraction of VN-collocations from text corpora: A feasibility study for German. In *Very Large Corpora: Academic and Industrial Perspectives*.

Bricken, T.; Mishra-Sharma, S.; Marcus, J.; Jermyn, A.; Olah, C.; Rivoire, K.; and Henighan, T. 2024. Stage-wise model diffing. *Transformer Circuits Thread*.

Bricken, T.; Templeton, A.; Batson, J.; Chen, B.; Jermyn, A.; Conerly, T.; Turner, N.; Anil, C.; Denison, C.; Askell, A.; et al. 2023. Towards monosemanticity: Decomposing language models with dictionary learning, 2023. *URL https://transformer-circuits.pub/2023/monosemantic-features/index. html*, 9.

Brown, B.; Juravsky, J.; Ehrlich, R.; Clark, R.; Le, Q. V.; Ré, C.; and Mirhoseini, A. 2024. Large language monkeys: Scaling inference compute with repeated sampling. *arXiv preprint arXiv:2407.21787*.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *NeurIPS*, 33: 1877–1901.

Cao, Y.; Zhang, T.; Cao, B.; Yin, Z.; Lin, L.; Ma, F.; and Chen, J. 2024. Personalized Steering of Large Language Models: Versatile Steering Vectors Through Bi-directional Preference Optimization. *arXiv preprint arXiv:2406.00045*.

Chanin, D.; Wilken-Smith, J.; Dulka, T.; Bhatnagar, H.; and Bloom, J. I. 2024. A is for Absorption: Studying Feature Splitting and Absorption in Sparse Autoencoders. In *Interpretable AI: Past, Present and Future*.

Chen, M.; Tworek, J.; Jun, H.; Yuan, Q.; Pinto, H. P. D. O.; Kaplan, J.; Edwards, H.; Burda, Y.; Joseph, N.; Brockman, G.; et al. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.

Chen, N.; Wu, N.; Liang, S.; Gong, M.; Shou, L.; Zhang, D.; and Li, J. 2023. Beyond surface: Probing llama across scales and layers. *arXiv preprint arXiv:2312.04333*.

Chinn, C. A.; and Anderson, R. C. 1998. The Structure of Discussions that Promote Reasoning. *Teachers College Record*, 100(2): 315–368.

Conerly, T.; Templeton, A.; Bricken, T.; Marcus, J.; and Henighan, T. 2024. Update on how we train SAEs. *Transformer Circuits Thread*.

Cywiński, B.; and Deja, K. 2025. SAeUron: Interpretable Concept Unlearning in Diffusion Models with Sparse Autoencoders. *arXiv preprint arXiv:2501.18052*.

Durmus, E.; Tamkin, A.; Clark, J.; Wei, J.; Marcus, J.; Batson, J.; Handa, K.; Lovitt, L.; Tong, M.; McCain, M.; et al. 2024. Evaluating feature steering: A case study in mitigating social biases, 2024. *URL https://anthropic. com/research/evaluating-feature-steering*.

Elhage, N.; Hume, T.; Olsson, C.; Schiefer, N.; Henighan, T.; Kravec, S.; Hatfield-Dodds, Z.; Lasenby, R.; Drain, D.; Chen, C.; et al. 2022. Toy models of superposition. *arXiv preprint arXiv:2209.10652*.

Gao, L.; Dupré la Tour, T.; Tillman, H.; Goh, G.; Troll, R.; Radford, A.; Sutskever, I.; Leike, J.; and Wu, J. 2024a. Scaling and evaluating sparse autoencoders. *arXiv preprint arXiv:2406.04093*.

Gao, L.; la Tour, T. D.; Tillman, H.; Goh, G.; Troll, R.; Radford, A.; Sutskever, I.; Leike, J.; and Wu, J. 2024b. Scaling and evaluating sparse autoencoders. arXiv:2406.04093.

Gerns, P.; and Mortimore, L. 2025. Towards exploratory talk in secondary-school CLIL: An empirical study of the cognitive discourse function 'explore'. *Language Teaching Research*.

Grattafiori, A.; Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Vaughan, A.; et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Guo, D.; Yang, D.; Zhang, H.; Song, J.; Zhang, R.; Xu, R.; Zhu, Q.; Ma, S.; Wang, P.; Bi, X.; et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Hendrycks, D.; Burns, C.; Kadavath, S.; Arora, A.; Basart, S.; Tang, E.; Song, D.; and Steinhardt, J. 2021. Measuring Mathematical Problem Solving With the MATH Dataset. arXiv:2103.03874.

Huben, R.; Cunningham, H.; Smith, L. R.; Ewart, A.; and Sharkey, L. 2024. Sparse Autoencoders Find Highly Interpretable Features in Language Models. In *ICLR*.

Jiang, Y.; Rajendran, G.; Ravikumar, P.; Aragam, B.; and Veitch, V. 2024. On the origins of linear representations in large language models. *arXiv preprint arXiv:2403.03867*.

Jin, M.; Yu, Q.; Huang, J.; Zeng, Q.; Wang, Z.; Hua, W.; Zhao, H.; Mei, K.; Meng, Y.; Ding, K.; et al. 2025. Exploring concept depth: How large language models acquire knowledge and concept at different layers? In *COLING*, 558–573.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kojima, T.; Gu, S. S.; Reid, M.; Matsuo, Y.; and Iwasawa, Y. 2022. Large language models are zero-shot reasoners. *NeurIPS*, 35: 22199–22213.

Kuznetsov, K.; Kushnareva, L.; Druzhinina, P.; Razzhigaev, A.; Voznyuk, A.; Piontkovskaya, I.; Burnaev, E.; and Barannikov, S. 2025. Feature-Level Insights into Artificial Text Detection with Sparse Autoencoders. *arXiv preprint arXiv:2503.03601*.

Leask, P.; Bussmann, B.; Pearce, M. T.; Bloom, J. I.; Tigges, C.; Al Moubayed, N.; Sharkey, L.; and Nanda, N. 2025. Sparse Autoencoders Do Not Find Canonical Units of Analysis. In *ICLR*.

Lieberum, T.; Rajamanoharan, S.; Conmy, A.; Smith, L.; Sonnerat, N.; Varma, V.; Kramár, J.; Dragan, A.; Shah, R.; and Nanda, N. 2024. Gemma scope: Open sparse autoencoders everywhere all at once on gemma 2. *arXiv preprint arXiv:2408.05147*.

Lindsey, J.; Templeton, A.; Marcus, J.; Conerly, T.; Batson, J.; and Olah, C. 2024. Sparse crosscoders for cross-layer features and model diffing. *Transformer Circuits Thread*.

MAA. 2024. AIME. https://artofproblemsolving.com/wiki/index.php?title=AIME_Problems_and_Solutions.

Madaan, A.; Tandon, N.; Gupta, P.; Hallinan, S.; Gao, L.; Wiegreffe, S.; Alon, U.; Dziri, N.; Prabhumoye, S.; Yang, Y.; et al. 2023. Self-refine: Iterative refinement with self-feedback. *NeurIPS*, 36: 46534–46594.

Marks, S.; Rager, C.; Michaud, E. J.; Belinkov, Y.; Bau, D.; and Mueller, A. 2024. Sparse feature circuits: Discovering and editing interpretable causal graphs in language models. *arXiv preprint arXiv:2403.19647*.

Meng, K.; Bau, D.; Andonian, A.; and Belinkov, Y. 2022. Locating and Editing Factual Associations in GPT. In *Advances in Neural Information Processing Systems*.

Michel, J.-B.; Shen, Y. K.; Aiden, A. P.; Veres, A.; Gray, M. K.; Team, G. B.; Pickett, J. P.; Hoiberg, D.; Clancy, D.; Norvig, P.; et al. 2011. Quantitative analysis of culture using millions of digitized books. *science*, 331(6014): 176–182.

Mihalcea, R.; and Tarau, P. 2004. Textrank: Bringing order into text. In *EMNLP*, 404–411.

Mikolov, T.; Chen, K.; Corrado, G.; and Dean, J. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Nanda, N.; Lee, A.; and Wattenberg, M. 2023. Emergent Linear Representations in World Models of Self-Supervised Sequence Models. In *Proceedings of the 6th BlackboxNLP Workshop*, 16–30.

Nostalgebraist. 2020. interpreting GPT: the logit lens. https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens. Accessed: 2025-12-05.

OpenAI. 2024. Learning to Reason with LLMs. https://openai.com/index/learning-to-reason-with-llms/.

Park, K.; Choe, Y. J.; and Veitch, V. 2023. The linear representation hypothesis and the geometry of large language models. *arXiv preprint arXiv:2311.03658*.

Paulo, G.; Mallen, A.; Juang, C.; and Belrose, N. 2024. Automatically interpreting millions of features in large language models. *arXiv preprint arXiv:2410.13928*.

Rayson, P.; and Garside, R. 2000. Comparing corpora using frequency profiling. In *Proceedings of the Workshop on Comparing Corpora - Volume 9*, WCC '00, 1–6. USA.

Rein, D.; Hou, B. L.; Stickland, A. C.; Petty, J.; Pang, R. Y.; Dirani, J.; Michael, J.; and Bowman, S. R. 2023. GPQA: A Graduate-Level Google-Proof Q&A Benchmark. arXiv:2311.12022.

Shao, Z.; Wang, P.; Zhu, Q.; Xu, R.; Song, J.; Bi, X.; Zhang, H.; Zhang, M.; Li, Y.; Wu, Y.; et al. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Shi, W.; Li, S.; Liang, T.; Wan, M.; Ma, G.; Wang, X.; and He, X. 2025. Route Sparse Autoencoder to Interpret Large Language Models. *arXiv preprint arXiv:2503.08200*.

Shinn, N.; Cassano, F.; Gopinath, A.; Narasimhan, K.; and Yao, S. 2023. Reflexion: Language agents with verbal reinforcement learning. *NeurIPS*, 36: 8634–8652.

Shu, D.; Wu, X.; Zhao, H.; Rai, D.; Yao, Z.; Liu, N.; and Du, M. 2025. A Survey on Sparse Autoencoders: Interpreting the Internal Mechanisms of Large Language Models. *arXiv preprint arXiv:2503.05613*.

Simonyan, K.; Vedaldi, A.; and Zisserman, A. 2014. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. In *ICLR*.

Snell, C.; Lee, J.; Xu, K.; and Kumar, A. 2024. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*.

Soboleva, D.; Al-Khateeb, F.; Myers, R.; Steeves, J. R.; Hestness, J.; and Dey, N. 2023. SlimPajama: A 627B token cleaned and deduplicated version of RedPajama. https://www.cerebras.net/blog/slimpajama-a-627b-token-cleaned-and-deduplicated-version-of-redpajama.

Soo, S.; Teng, W.; and Balaganesh, C. 2025. Steering Large Language Models with Feature Guided Activation Additions. *arXiv preprint arXiv:2501.09929*.

Team, Q. 2024. Qwq: Reflect deeply on the boundaries of the unknown. *Hugging Face*.

Templeton, A. 2024. *Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet*. Anthropic.

Thoughts, O. 2025. Open Thoughts. https://open-thoughts.ai. Accessed: 2025-12-05.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *NeurIPS*, 5998–6008.

Wang, X.; Wei, J.; Schuurmans, D.; Le, Q.; Chi, E.; Narang, S.; Chowdhery, A.; and Zhou, D. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *NeurIPS*, 35: 24824–24837.

Yan, H.; Xiang, Y.; Chen, G.; Wang, Y.; Gui, L.; and He, Y. 2024. Encourage or Inhibit Monosemanticity? Revisiting Monosemanticity from a Feature Decorrelation Perspective. *arXiv:2406.17969*.

Zhang, M.; Li, X.; Yue, S.; and Yang, L. 2020. An empirical study of TextRank for keyword extraction. *IEEE access*, 8: 178849–178858.

Zheng, L.; Chiang, W.-L.; Sheng, Y.; Li, T.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Li, Z.; Lin, Z.; Xing, E. P.; et al. 2023. Lmsys-chat-1m: A large-scale real-world llm conversation dataset. *arXiv preprint arXiv:2309.11998*.