The Reasoning-Memorization Interplay in Language Models Is Mediated by a Single Direction

Anonymous ACL submission

Abstract

Large language models (LLMs) excel on a variety of reasoning benchmarks, but previous studies suggest they sometimes struggle to generalize to unseen questions, potentially due to over-reliance on memorized training examples. However, the precise conditions under which LLMs switch between reasoning and memorization during text generation remain unclear. In this work, we provide a mechanistic understanding of LLMs' reasoning-memorization dynamics by identifying a set of linear features in the model's residual stream that govern the balance between genuine reasoning and memory recall. These features not only distinguish reasoning tasks from memory-intensive ones but can also be manipulated to causally influence model performance on reasoning tasks. Additionally, we show that intervening in these reasoning features helps the model more accurately activate the most relevant problem-solving capabilities during answer generation. Our findings offer new insights into the underlying mechanisms of reasoning and memory in LLMs and pave the way for the development of more robust and interpretable generative AI systems.[1]

1 Introduction

Large language models (LLMs) have demonstrated impressive capabilities in tackling complex reasoning tasks (Roziere et al., 2023; OpenAI, 2024; Guo et al., 2025). However, these models sometimes struggle with more straightforward reasoning problems, particularly when faced with questions that differ significantly from those encountered during training (Dziri et al., 2024; Hu et al., 2024; Xie et al., 2024). This generalization gap between LLMs and human reasoning has led to the hypothesis that these models are essentially "reasoning parrots" (Zečević et al., 2023), relying heavily on

[1] Our code and data have been uploaded to the submission system, and will be open-sourced upon acceptance.

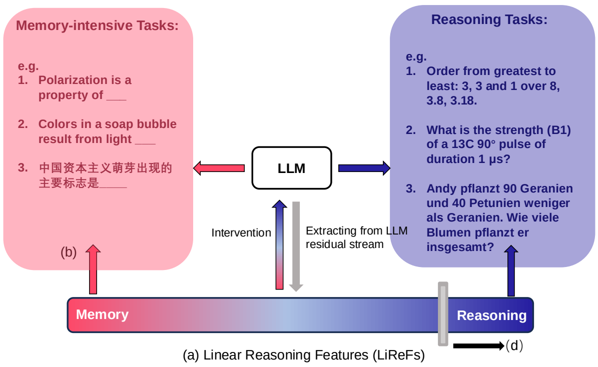

Figure 1: Main findings of our study: (a) There exists a set of linear features (LiReFs) in the LLM residual stream that drives the model to switch between reasoning and memorization modes with different levels of generalizability. (b) LiReFs generally explain model reasoning capability across various knowledge domains and languages. (c) Model activation values along LiReFs correlate strongly with model generalizability on reasoning tasks. (d) Intervening LiReFs during inference time can further improve the model reasoning performance and generalizability.

memorization of text patterns found in their pretraining datasets (Carlini et al., 2022; Tang et al., 2023; Shi et al., 2024), rather than engaging in a rigorous, procedural reasoning process to solve problems (Wei et al., 2022; Kojima et al., 2022; Yao et al., 2023). Understanding the interplay between reasoning and memorization in LLMs is essential, not only for advancing our understanding of these models but also for developing more reliable, language-based reasoning systems in the future (Lanham et al., 2023; Oren et al., 2023; Turpin et al., 2024).

In the context of LLM reasoning, researchers often conceptualize memorization as the inability to generalize from familiar problems to their systematically modified counterparts. In this view, reasoning and memorization are two extremes on the spectrum of model generalizability. To investigate this, synthetic reasoning benchmarks are designed, and memorization is assessed by measuring changes in model performance across various setups (Dziri et al., 2024; Xie et al., 2024; Ye et al., 2024). Another line of research focuses on the internal mechanisms of LLMs, identifying specific components or circuits responsible for tasks like arithmetic (Hou et al., 2023; Stolfo et al., 2023a) and commonsense reasoning (Geva et al., 2023; Yang et al., 2024; Biran et al., 2024). However, these studies primarily analyze model outputs or hidden representations when dealing with carefully crafted synthetic reasoning problems, limiting the generalizability of their findings.

In this paper, we explore the reasoning-memorization dynamic of LLMs from a mechanistic perspective. Recent interpretability research has demonstrated that LLMs encode interpretable semantic features (Elhage et al., 2022; Park et al., 2024)—such as safety (Arditi et al., 2024; Yu et al., 2024), truth (Marks and Tegmark, 2023; Li et al., 2024), sentiment (Tigges et al., 2023), and language (Bricken et al., 2023)—as linear directions within their activation space. We hypothesize that there is a similar linear feature, which, when activated, enables the model to solve reasoning tasks through systematic generalization. When this feature is not activated, the model remains in a "memorization mode," exhibiting low generalizability when addressing variations of familiar reasoning problems.

To examine our hypotheses, we apply methods from linear semantic feature analysis (Burns et al., 2023; Rimsky et al., 2024) and identify a set of Linear Reasoning Features (LiReFs) in the residual streams of LLMs. As shown in Figure 1, LiReFs can be extracted by contrasting the hidden representations of reasoning-intensive versus memory-intensive questions. This contrast allows the two types of questions to be linearly separated in the model's activation space. Furthermore, we demonstrate via causal analysis (Stickland et al., 2024; Hong et al., 2024) that by enhancing the LiReFs during inference, we can shift the model into a "thinking mode" with strong generalizability in applying reasoning rules or patterns. We show via extensive experiments on four different LLMs across six datasets that the same set of reasoning features explain and mediate model reasoning ability across various knowledge domains and languages, suggesting a general control mechanism of switching between reasoning and memorization during model inference.

The main contributions of our work can be summarized as follows:

- We show that LLM reasoning capability is mediated by a set of linear features (LiReFs) in its activation space. Such features govern model generalizability in solving various reasoning tasks including math, logical, and scientific questions (Section 3).
- We casually validate the functionality of our discovered reasoning features by showing that LLM reasoning generalizability can be enhanced by intervening LiReFs at inference time (Section 4.1).
- We show via case analyses that mediating LiReFs during inference time reduces LLM reasoning errors and misapplication of model reasoning or memorization ability. (Section 4.2).

2 Related work

Memorization in LLMs Memorization in LLMs has been defined in various ways. In the context of privacy and copyright, memorization is often described as the model's verbatim reproduction of training data during generation (Carlini et al., 2022; Biderman et al., 2023; Huang et al., 2024). Alternatively, some define memorization as the counterfactual effect of omitting specific training data on model predictions (Zhang et al., 2023; Hu et al., 2024), reflecting memorization of rare, specific examples. In reasoning tasks, memorization is often seen as poor generalizability to questions outside the training data, as evidenced by studies on work sequence reversal (McCoy et al., 2023) and alphabet shifting (Prabhakar et al., 2024), which show degraded performance on infrequent patterns. Other studies observe performance degradation from controlled perturbations of input questions (Wu et al., 2024; Xie et al., 2024). In this paper, we adopt memorization as poor reasoning generalizability and propose a novel mechanistic interpretation of the reasoning-memorization dynamic during model inference.

Understanding LLM reasoning Prior research has sought to distinguish reasoning from memorization, investigating whether LLMs genuinely infer new conclusions or merely reconstruct patterns from pretraining data. Studies suggest that

LLMs undergo structured multi-step reasoning processes, transitioning through distinct reasoning stages that follow an ordered sequence of knowledge retrieval and rule-based processing (Hou et al., 2023). Similarly, extended training beyond overfitting (grokking) has been shown to lead to the emergence of reasoning circuits, indicating that reasoning is a learned and structured capability (Power et al., 2022; Liu et al., 2022; Nanda et al., 2023; Wang et al., 2024a). Further studies on mathematical reasoning confirm that LLMs compute necessary information rather than memorizing templates, with reasoning computations leaving identifiable traces in model activations, particularly in the residual stream (Ye et al., 2024; Stolfo et al., 2023b). Additionally, attention heads have been shown to play a key role in both knowledge recall and latent reasoning, suggesting that these processes are distinct yet interconnected (Zheng et al., 2024).

Linear semantic features Recent advances in model interpretability have revealed that language models encode various semantic concepts as linear directions in their activation space (Park et al., 2024). These linear semantic features have been discovered by contrasting inputs that differ primarily in the targe semantic dimension (Marks and Tegmark, 2023). Once identified, these linear features can be manipulated to control model behavior, enabling targeted interventions during the generation process (Rimsky et al., 2024; Stickland et al., 2024). Our work extends this line of study by identifying linear features that mediate the model's ability to switch between genuine reasoning and memory recall.

3 Linear reasoning features (LiReFs)

3.1 Background

Transformers A decoder-only transformer language model (Vaswani et al., 2017) $\mathcal{M}$ maps an input sequence of tokens $x = [x_1, ..., x_T]$ into a probability distribution over the vocabulary for next-token prediction. Within the transformer, the i-th token x_i is represented as a series of hidden states $\mathbf{h}^{(l)}(x_i)$. Within each layer $l \in [L]$, two modules compute updates that are added to the layer input $\mathbf{h}^{(l-1)}(x_i)$: (1) a **multi-head self-attention module** outputs $\mathbf{a}^{(l)}(x_i)$, and a **multi-layer perceptron (MLP)** outputs $\mathbf{m}^{(l)}(x_i)$. Putting together, the hidden representation $\mathbf{h}^{(l)}(x_i)$ is computed as [2]:

$$\mathbf{h}^{(l)}(x_i) = \mathbf{h}^{(l-1)}(x_i) + \mathbf{a}^{(l)}(x_i) + \mathbf{m}^{(l)}(x_i) \quad (1)$$

Following Elhage et al. (2021), we call each $\mathbf{h}^{(l)}(x_i)$ the *residual stream activation* of x_i at layer l. We focus on the residual stream of the last token x_T of the user turn, as the point when the model is going to generate the first answer token, denoted as $\mathbf{H}(x) = \{\mathbf{h}^{(l)}(x_T)\}_{l=1}^{L}$.

Reasoning feature extraction We follow the linear feature hypothesis and postulate that the reasoning capability of LLMs is mediated by a single direction in the residual stream, and that by steering this direction, it is possible to control model interplay between reasoning and memorization. We compute the *linear reasoning features (LiReFs)* using the *difference-in-means* technique, which effectively disentangles key feature information as demonstrated by previous work (Marks and Tegmark, 2023; Rimsky et al., 2024). Specifically, given a collection of *reasoning-intensive questions* $x \in \mathcal{D}_{\textbf{Reasoning}}$ (e.g. "What is the answer of $(5+2)*3$?") and another set of *memory-intensive questions* $x \in \mathcal{D}_{\textbf{Memory}}$ (e.g. "What is the capital city of the USA?"), we calculate the difference between the model's mean last-token residual stream activations when running on two categories of input questions:

$$\mathbf{r}^{(l)} = \frac{\sum\limits_{x \in \mathcal{D}_{\textbf{Reasoning}}} \mathbf{h}^{(l)}(x)}{|\mathcal{D}_{\textbf{Reasoning}}|} - \frac{\sum\limits_{x \in \mathcal{D}_{\textbf{Memory}}} \mathbf{h}^{(l)}(x)}{|\mathcal{D}_{\textbf{Memory}}|} \quad (2)$$

The specific construction details of $\mathcal{D}_{\textbf{Memory}}$ and $\mathcal{D}_{\textbf{Reasoning}}$ are provided in Section 3.2.

Reasoning feature intervention Given a difference-in-means vector $\mathbf{r}^{(l)}$ extracted from layer l, we can modulate the strength of the corresponding reasoning feature via simple linear interventions. Specifically, we can perform *reasoning feature addition* by adding the difference-in-means vector to the activations of an input question to shift it closer to the mean activation of typical reasoning-intensive questions, thereby unlocking model reasoning capability:

$$\mathbf{h}'^{(l)}(x) \leftarrow \mathbf{h}^{(l)}(x) + \alpha * \mathbf{r}^{(l)} \quad (3)$$

Similarly, one can perform *reasoning feature ablation* by erasing the component along $\hat{\mathbf{r}}^{(l)}$ for

[2] Here, we omit some details such as positional encoding and layer normalization for brevity.

every residual stream activation $\mathbf{h}^{(l)}(x)$:

$$\mathbf{h}'^{(l)}(x) \leftarrow \mathbf{h}^{(l)}(x) - \hat{\mathbf{r}}\hat{\mathbf{r}}^T\mathbf{h}^{(l)}(x) \tag{4}$$

where $\hat{\mathbf{r}} = \mathbf{r}^{(l)}/||\mathbf{r}^{(l)}||$ is a unit vector encoding the reasoning feature direction, and $\mathbf{h}^{(l)}(x) - \hat{\mathbf{r}}\hat{\mathbf{r}}^T\mathbf{h}^{(l)}(x)$ is projection that zeroes out the value along the reasoning direction.

3.2 Datasets and Models

Datasets We curate our dataset for LiReF extraction and analysis using the following existing question answering benchmarks: 1) MMLU-Pro (Wang et al., 2024b), which is a comprehensive QA benchmark covering a wide range of subjects, including STEM, humanities and social sciences fields; 2) the GSM-8K math reasoning dataset (Cobbe et al., 2021) and its multilingual counterpart MGSM (Shi et al., 2022); 3) the PopQA factual knowledge QA dataset (Mallen et al., 2023), and 4) the humanity sections of the C-Eval Chinese benchmark (Huang et al., 2023). A detailed description of each dataset can be found in §B.

To categorize QA questions into the contrastive reasoning-intensive and memory-intensive subsets, we employ LLM-as-a-judge (Zheng et al., 2023) by asking GPT-4o (OpenAI et al., 2024) to assign a score between 0 and 1 to each question in MMLU-Pro, where a score closer to 1 indicates a reasoning-intensive question, and a score closer to 0 suggests a memory-intensive one. A score around 0.5 indicates that both reasoning and memory recall may be involved [3]. Next, we classified questions with scores above 0.5 as MMLU-Pro-R (Reasoning Part) and placed them in $\mathcal{D}_{\textbf{Reasoning}}$, while questions with scores less than or equal to 0.5 were classified as MMLU-Pro-M (Memory Part) and placed in $\mathcal{D}_{\textbf{Memory}}$. For the other benchmarks, we assign GSM8K and MGSM into $\mathcal{D}_{\textbf{Reasoning}}$, and put PopQA and C-Eval Chinese into $\mathcal{D}_{\textbf{Memory}}$.

Models We study LiReF by analyzing a diverse collection of representative and influential base models, as long as their instruction-tuned variants: LLaMA3-8B (base, instruct) (Grattafiori et al., 2024), Gemma2-9B (base, instruct) (Team et al., 2024), Mistrial-7B-v0.3 (base, instruct) (Jiang et al., 2023), and OLMo2-7B (base, instruct) (OLMo et al., 2025).

3.3 Analysis results

Figure 2 shows the 2-dimensional Principal Component Analysis (PCA) visualization of the last tokens representations across different model layers and six datasets in $\mathcal{D}_{\textbf{Memory}}$ and $\mathcal{D}_{\textbf{Reasoning}}$, where hidden representations are taken from a specific middle layer of each model. [4] Additional PCA results for other layers of the models are provided in Appendix C. We observe that the representations of questions in $\mathcal{D}_{\textbf{Memory}}$ and $\mathcal{D}_{\textbf{Reasoning}}$ can be linearly separated by the reasoning features, which are computed as the difference vector between centroids of the two representation categories (the blue arrows).

Robustness of LiReF extraction We also validate that our extracted LiReFs indeed capture model reasoning capability, as opposed to some superficial lexical patterns that distinguish two question categories. As suggested by Figure 2, for each model, the same LiReF separates every contrastive pair of problem subsets in $\mathcal{D}_{\textbf{Reasoning}}$ and $\mathcal{D}_{\textbf{Memory}}$, regardless of the task format (e.g., multiple choice and the open-ended generation), domain (e.g., physics, chemistry and math), or language (e.g., English and Chinese). Moreover, we provide in Appendix C more fine-grained PCA visualizations of questions from various subject domains in MMLU-Pro, suggesting that even for questions from disparate disciplines (e.g., physics vs. history), as long as both of their solutions require strong reasoning capability, their hidden representations shall fall into the same reasoning subspace as determined by the LiReF.

To quantitatively measure the relation between LiReF and the reasoning capability required for answering each question, we compute the layerwise cosine similarity between the last question token representation of each question and the corresponding LiReF, as shown in Figure 3. For each LLM, we also replicate the same analyses for its pre-trained base version before instruction fine-tuning. A positive cosine similarity suggests a positive activation value along LiReF and vice versa. We observe that for all eight models, questions in $\mathcal{D}_{\textbf{Reasoning}}$ mostly activate the reasoning features positively, while questions in $\mathcal{D}_{\textbf{Memory}}$ mostly have negative

[3] The prompt used is provided in §A.

[4] Figure 10 in the Appendix C shows that the top one principal component already captures most of the mean difference (see Equation 2) between the activations in $\mathcal{D}_{\textbf{Memory}}$ and $\mathcal{D}_{\textbf{Reasoning}}$.

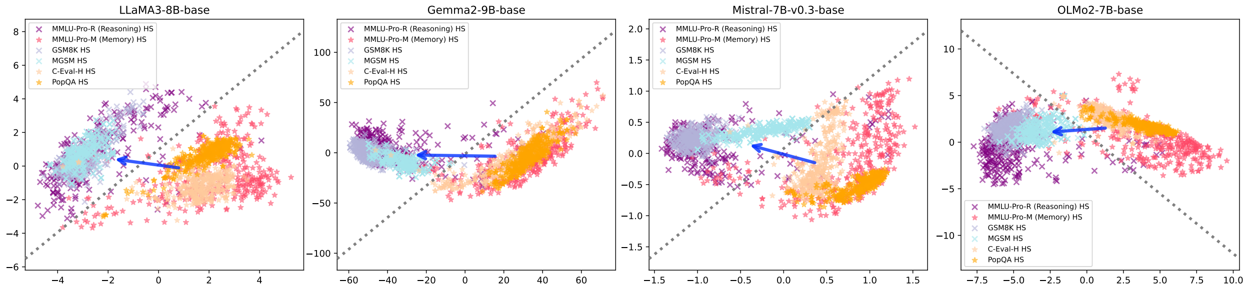

Figure 2: Visualization of the hidden states of four base models using 2-dimensional PCA. For each model, we plot six groups of points across several datasets. We observe that: (1) For all four models, questions defined as Reasoning-required and those defined as Memory-required can be naturally distinguished into two distinct groups, as shown by the boundary (grey dashed line) fitted via logistic regression, with the blue arrows showing the approximate direction of the Linear Reasoning Features. (2) In the extracted dimensions, the influence of task domain and language within the same category on the distribution is not significant, and data requiring the same capability naturally cluster together in the same region.

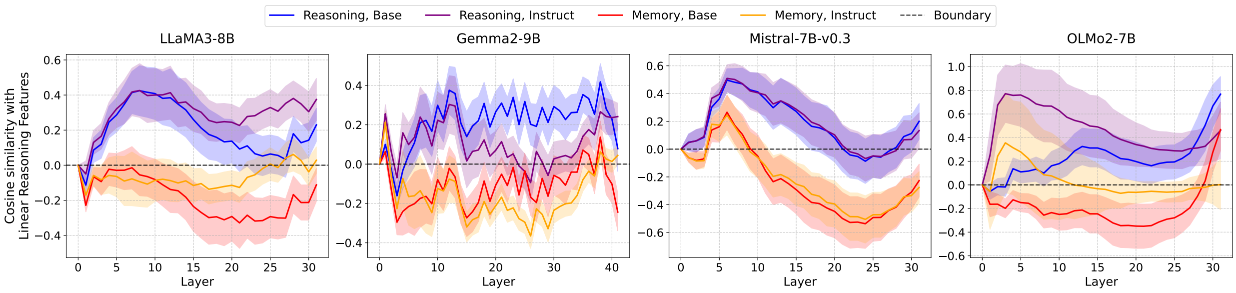

Figure 3: Layerwise cosine similarity between the last token residual stream activations and the extracted Linear Reasoning Features (LiReFs) in four base models and their corresponding instruction-tuned variants.

LiReF activations, especially in the middle layers. Furthermore, on 3 out of 4 LLM families (LLaMA3-8B, Gemma2-9B, and Mistral-7B-v0.3), the layerwise cosine similarity profiles between the base and instruction-tuned models are highly consistent with each other, suggesting that LLMs may have developed linear reasoning features to mediate its emergent reasoning capability during pre-training rather than post-training.

3.4 The gradient nature of reasoning-memorization interplay

As observed in Figure 2, questions in $\mathcal{D}_{\textbf{Memory}}$ and $\mathcal{D}_{\textbf{Reasoning}}$ tend to have significantly negative and positive activations along LiReFs, respectively. This raises the question: what types of questions fall near the reasoning-memorization boundary (i.e., those with near-zero LiReF activation values)? Do these problems require both memory and reasoning abilities to solve? We investigate this question through the following experiments.

Figure 4 shows the relation between GPT-4o-assigned reasoning scores for each question in MMLU-Pro, as discussed in Section 3.2, versus the LiReF projection value $\hat{\mathbf{r}}^T\mathbf{h}^{(l)}(x)$ of its residual stream representation $\mathbf{h}^{(l)}(x)$ by LLaMA3-8B-base and Mistral-7B-v0.3 models. We observe that as problems receive higher reasoning scores assigned by GPT-4o, they tend to have larger activation values along the LiReF direction. This correlation is notably strong across both models, with Spearman correlation coefficients of 0.840 for LLaMA3-8B-base and 0.752 for Mistral-7B-v0.3-

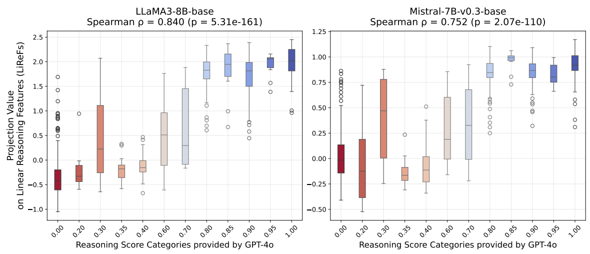

Figure 4: Strong correlation between Projection Values on the Linear Reasoning Features (LiReFs) direction and the Reasoning Score provided by GPT-4o, with Spearman coefficients of 0.840 (LLaMA3-8B-base) and 0.752 (Mistral-7B-v0.3-base). The LiReFs projections exhibit a spectrum-like distribution, where continuous increases in Reasoning Scores correspond to progressively rising Projection Values along the LiReFs direction.

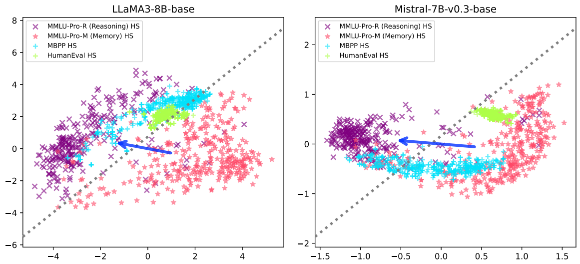

Figure 5: Visualization of the hidden states of two base models on the datasets of MBPP, HumanEval, MMLU-Pro-M and MMLU-Pro-R using 2-dimensional PCA. The hidden states of coding tasks, which involve both reasoning and memory recall, are positioned around the boundary (grey dashed line) fitted via logistic regression.

base. These findings suggest that problems with near-zero LiReF activations likely involve both memory and reasoning capabilities.

To further validate our results, we conducted additional PCA experiments on the Coding tasks - which have been identified by numerous studies as a representative task type requiring both memory and reasoning capabilities in LLMs (Zhao et al., 2025; Chen et al., 2024). The results are shown in Figure 5, where we observe that the residual stream activations of two Coding tasks, MBPP (Austin et al., 2021) and HumanEval (Chen et al., 2021), are both positioned near the boundary. This further supports our finding that data points situated between the two extremes represent task types that engage both memory and reasoning abilities in LLMs.

4 Causal validation of LiReFs

4.1 Inference-time LiReF intervention

In this section, we conduct experiments where we manually intervene in the residual stream activations during inference time. By adjusting the intensity of linear reasoning features in model residual streams, we examine how model performance on both memory-intensive and reasoning-intensive tasks will change.

In particular, for all tokens of each question, we modify their residual stream representations in a specific layer by adding an intervention vector along the LiReF direction, as suggested in Equation 3. To enhance the most relevant model capability, we adopt negative values of α for $\mathcal{D}_{\textbf{Memory}}$, and positive α values for $\mathcal{D}_{\textbf{Reasoning}}$. After carefully tuning α on validation sets, we ask each model to generate answers for questions in $\mathcal{D}_{\textbf{Memory}}$ and $\mathcal{D}_{\textbf{Reasoning}}$, and measure its performance change under inference-time LiReF intervention. More details about the experimental setup, including the validation-test set splits, hyperparameter selection criteria and inference settings can be found in Appendix D.

The main results are shown in Table 1. We observe that intervening LiReFs during inference time effectively improves the performance of four LLMs on both memory-intensive and reasoning-intensive tasks. Moreover, the improvements remain consistent across different task types, domains, and languages, further supporting our claim that the reasoning features in LLM residual streams capture general reasoning capability. In the next section, we will present specific cases to illustrate how reasoning feature intervention improves model performance by reducing reasoning step errors and correcting the misapplication of model abilities.

4.2 Cases Study

In the PCA analyses presented in Section 3.3, we observed certain sample cases that, although labeled as reasoning-intensive by GPT-4o or by the task name, have negative LiReF activations on the memorization subspace. Similarly, some cases that were labeled as memory-intensive instead fall into the reasoning subspace with positive-valued LiReFs. In this section, we analyze these cases and also conduct LiReF intervention experiments, aiming to correct any potential reasoning errors or unfaithful reasoning steps.

Firstly, we collect questions in MMLU-Pro whose reasoning label contradicts the actual feature subspace in which they are positioned. (e.g., cases whose GPT-4o-assigned reasoning score is much less than 0.5, but have a positive-valued LiReF activation), and evaluate LLaMA3-8B-base on them to identify a subset of questions where the model provides incorrect answers. Then we obtained a subset of 184 cases in total. Next, we perform inference-time LiReF intervention on these examples following the same settings in Section 4.1, and compare their accuracy and actual outputs before and after the intervention. We found that, by shifting LiReF activation to have the sign that is consistent with GPT-4o-assigned reasoning score, model accuracy on this subset jumps from 0 to 0.21. Table 2 presents some exemplar questions in our analyses, together with model answers before and after LiReF intervention. These results suggest that

	Memory-Intensive Dataset			Reasoning Dataset		
Base model	MMLU-Pro-M	PopQA	C-Eval-H	MMLU-Pro-R	GSM-8k	MGSM
LLaMA3-8B-base	41.1 / **48.3** $\uparrow$7.2	33.4 / **35.6** $\uparrow$2.2	45.2 / **47.4** $\uparrow$2.2	24.2 / **33.5** $\uparrow$9.3	49.0 / **53.1** $\uparrow$4.1	28.5 / **34.6** $\uparrow$6.1
Gemma2-9B-base	37.5 / **50.1** $\uparrow$12.6	29.2 / **30.3** $\uparrow$1.1	52.1 / 52.1	29.2 / **44.7** $\uparrow$15.5	61.9 / **63.5** $\uparrow$1.6	45.8 / **47.0** $\uparrow$1.2
Mistral-7B-v0.3-base	37.8 / **43.6** $\uparrow$5.8	30.1 / **30.9** $\uparrow$0.8	38.2 / **44.0** $\uparrow$5.8	20.8 / **21.7** $\uparrow$0.9	35.1 / **36.2** $\uparrow$1.1	12.0 / 12.0
OLMo2-7B-base	19.4 / **25.0** $\uparrow$5.6	19.2 / **20.1** $\uparrow$0.9	26.0 / **28.9** $\uparrow$2.9	11.3 / **16.5** $\uparrow$5.2	11.5 / **12.3** $\uparrow$0.8	10.1 / **11.3** $\uparrow$1.2

Table 1: The performance of four base models on six benchmarks, before and after feature intervention. The results indicate that by shifting the residual stream of the reasoning-required or memory-required tasks further to the specific feature regions, overall task performance can be substantially enhanced.

LLM reasoning errors might not be due to a lack of relevant knowledge, but are caused by the insufficient activation of its acquired generalizable thinking capabilities, which can be alleviated through targeted inference-time intervention of reasoning features.

4.3 Reasoning Generalization Effects

In the previous experiments, we noticed that the features of certain questions from reasoning datasets lie in the memory subspace with negative LiReF activations. Therefore, we suspect that the models might have solved these reasoning questions through memorization (possibly due to training data contamination), rather than applying genuine reasoning capability that is generalizable under systematic input variation. To verify this hypothesis, we conduct additional features intervention experiments on GSM-Symbolic (Mirzadeh et al., 2025) in this section. GSM-Symbolic is a variant of GSM-8k. It selects 100 question templates from GSM-8k and then generates 50 different instances for each template by varying numerical conditions, results, and other factors. The resulting dataset contains 5,000 data points, making it ideal for a reliable evaluation of the model's reasoning generalization capabilities.

Figure 6 shows mean model accuracy on GSM-Symbolic, GSM-8k, and MMLU-Pro-M under inference-time LiReF intervention. We can see that as the intervention intensity α increases from 0, the performance of all four models on both GSM-8k and GSM-Symbolic rises consistently. On the other hand, as α decreases from 0, we observe that, compared to GSM-8k, GSM-Symbolic experiences a more significant performance drop with suppressed LiReFs. Notably, the performance gain and loss on GSM-Symbolic suggests that LiReF intervention is likely enhancing the genuine model reasoning capability that is generalizable, as opposed to case-based reasoning skills that rely more on memorization of particular training examples. Interestingly, we also observe that the performance drop on GSM-8K under LiReF suppression is less pronounced compared to GSM-Symbolic, and there is even a slight improvement with a moderate suppression when setting $\alpha = -0.05$. This implies that the model might have previously been exposed to GSM-8K due to data leakage, and therefore adopts a memory-intensive strategy to answer these questions. While for MMLU-Pro-M, performance improves steadily as α decreases, supporting our observation that reducing the projection values of the model's residual stream in the LiReFs direction enhances the model's ability to recall from memory.

5 Conclusion

In this study, we show that the reasoning and memorization abilities of LLMs across various domains and languages are driven by a shared set of linear reasoning features within their activation space. These features can be uncovered by contrasting the hidden representations of reasoning-intensive and memory-intensive queries. Furthermore, we demonstrate that intervening on these Linear Reasoning Features (LiReFs) during model inference improves both accuracy and robustness to perturbations when handling reasoning tasks. We also reveal that manipulating these reasoning features enables the model to more precisely activate the most relevant problem-solving capabilities during answer generation. Our findings offer a mechanistic understanding of how reasoning and memorization interact in LLMs, providing a foundation for future research aimed at building more capable and interpretable generative reasoning systems.

6 Limitations

Our work has several limitations. First, we only studied reasoning features in relatively small LLMs, while recent studies show that by scaling up both

Example Question	Answer before Intervention	Correct Answer after Intervention	Error Reason
Speed of data transmission in 4-G network of telecom is?	4G networks are designed to provide ultra-fast speeds, and with advancements in LTE-Advanced and carrier aggregation, they can reach speeds as high as 10 Gbps. The answer is **10 Gbps**.	The speed of data transmission in a 4G network typically ranges from 100 Mbps (megabits per second) to 1 Gbps (gigabit per second), depending on factors such as network congestion. The answer is **100 Mbps to 1 Gbps**.	Shallow Memory recall
A sound wave with frequency f travels through air at speed v. With what speed will a sound wave with frequency 4f travel through the air?	4f is four times the frequency of f, so the wavelength of 4f is one-fourth the wavelength of f. Since the speed of sound is the product of frequency and wavelength, the speed of 4f is one-fourth the speed of f. The answer is **v/4**.	The speed of sound in air is independent of the frequency of the sound wave. It only depends on factors like the medium and the temperature of the air. So, if a sound wave with frequency f travels through air at speed v, a sound wave with frequency 4f will also travel through the air at the same speed, v. The answer is **v**.	Overthinking

Table 2: Example questions and outputs before and after features intervention on the LLaMA3-8B-base model. They are both Memory-Intensive questions, but the features were incorrectly positioned in the Reasoning region. After shifting the Features to the appropriate region, the error was corrected.

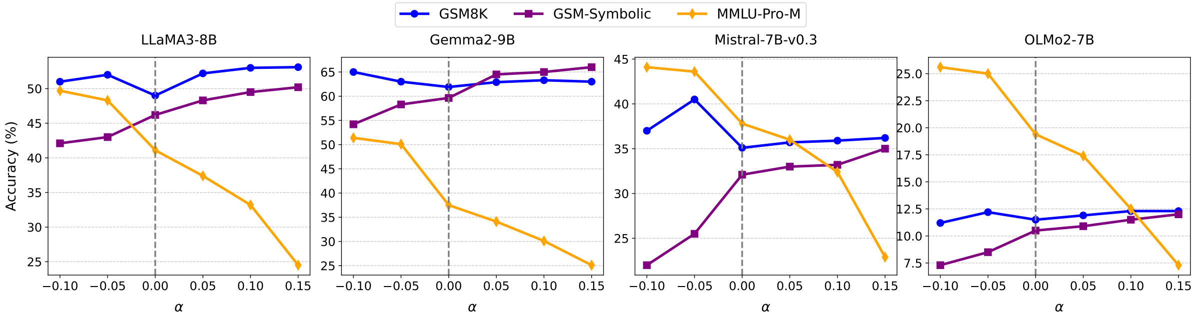

Figure 6: Performance of the four base models on the GSM-8k, GSM-Symbolic, and MMLU-Pro-M datasets, with varying hyperparameter α to control the intensity of feature intervention.

model size and inference-time computation, the reasoning capability of LLMs can be significantly improved (Hoffmann et al., 2022; OpenAI, 2024). Second, we have focused mostly on reasoning problems that can be addressed through short answers, while it remains unclear whether LiReFs can be utilized to enhance model's ability of performing deliberate reasoning via various prompt engineering techniques such as chain-of-thought (Wei et al., 2022), self-reflection (Shinn et al., 2024), and tree-of-thought (Yao et al., 2024). Third, we formulate memorization as performance inconsistency against reasoning question perturbation, while another line of LLM reasoning research has employed a different definition of *counterfactual memorization* – i.e., change of model answers on particular test questions after removing a similar example from training data (Zhang et al., 2023; Hu et al., 2024). Future work should investigate Whether perturbational and counterfactual memorization are mechanistically equivalent and, therefore, can be both mediated by LiReFs.

References

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. 2024. Refusal in language models is mediated by a single direction. *arXiv preprint arXiv:2406.11717*.

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, and Charles Sutton. 2021. Program synthesis with large language models. *Preprint*, arXiv:2108.07732.

Stella Biderman, USVSN Sai Prashanth, Lintang Sutawika, Hailey Schoelkopf, Quentin Anthony, Shivanshu Purohit, and Edward Raff. 2023. Emergent and predictable memorization in large language models. *Preprint*, arXiv:2304.11158.

Eden Biran, Daniela Gottesman, Sohee Yang, Mor Geva, and Amir Globerson. 2024. Hopping too late: Exploring the limitations of large language models on multi-hop queries. *arXiv preprint arXiv:2406.12775*.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. 2023. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*. Https://transformer-circuits.pub/2023/monosemantic-features/index.html.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. 2023. Discovering latent knowledge in language models without supervision. In *The Eleventh International Conference on Learning Representations*.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. 2022. Quantifying memorization across neural language models. In *The Eleventh International Conference on Learning Representations*.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. 2021. Evaluating large language models trained on code. *Preprint*, arXiv:2107.03374.

Xinyun Chen, Maxwell Lin, Nathanael Schärli, and Denny Zhou. 2024. Teaching large language models to self-debug. In *The Twelfth International Conference on Learning Representations*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *Preprint*, arXiv:2110.14168.

Nouha Dziri, Ximing Lu, Melanie Sclar, Xiang Lorraine Li, Liwei Jiang, Bill Yuchen Lin, Sean Welleck, Peter West, Chandra Bhagavatula, Ronan Le Bras, et al. 2024. Faith and fate: Limits of transformers on compositionality. *Advances in Neural Information Processing Systems*, 36.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, et al. 2022. Toy models of superposition. *arXiv preprint arXiv:2209.10652*.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. 2021. A mathematical framework for transformer circuits. `https://transformer-circuits.pub/2021/framework/index.html`. Published: Dec 22, 2021.

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. 2023. Dissecting recall of factual associations in auto-regressive language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12216–12235, Singapore. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh,

Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary

DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. 2022. Training compute-optimal large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, pages 30016–30030.

Yihuai Hong, Lei Yu, Shauli Ravfogel, Haiqin Yang, and Mor Geva. 2024. Intrinsic evaluation of unlearning using parametric knowledge traces. *arXiv preprint arXiv:2406.11614*.

Yifan Hou, Jiaoda Li, Yu Fei, Alessandro Stolfo, Wangchunshu Zhou, Guangtao Zeng, Antoine Bosselut, and Mrinmaya Sachan. 2023. Towards a mechanistic interpretation of multi-step reasoning capabilities of language models. *arXiv preprint arXiv:2310.14491*.

Yi Hu, Xiaojuan Tang, Haotong Yang, and Muhan Zhang. 2024. Case-based or rule-based: How do transformers do the math? *arXiv preprint arXiv:2402.17709*.

Jing Huang, Diyi Yang, and Christopher Potts. 2024. Demystifying verbatim memorization in large language models. *Preprint*, arXiv:2407.17817.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, jiayi lei, Yao Fu, Maosong Sun, and Junxian He. 2023. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, et al. 2023. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2024. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems*, 36.

Ziming Liu, Ouail Kitouni, Niklas S Nolte, Eric Michaud, Max Tegmark, and Mike Williams. 2022. Towards understanding grokking: An effective theory of representation learning. *Advances in Neural Information Processing Systems*, 35:34651–34663.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9802–9822, Toronto, Canada. Association for Computational Linguistics.

Samuel Marks and Max Tegmark. 2023. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. *arXiv preprint arXiv:2310.06824*.

R. Thomas McCoy, Shunyu Yao, Dan Friedman, Matthew Hardy, and Thomas L. Griffiths. 2023. Embers of autoregression: Understanding large language models through the problem they are trained to solve. *Preprint*, arXiv:2309.13638.

Seyed Iman Mirzadeh, Keivan Alizadeh, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, and Mehrdad Farajtabar. 2025. GSM-symbolic: Understanding the limitations of mathematical reasoning in large language models. In *The Thirteenth International Conference on Learning Representations*.

Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. 2023. Progress measures for grokking via mechanistic interpretability. In *The Eleventh International Conference on Learning Representations*.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, Michal Guerquin, Hamish Ivison, Pang Wei Koh, Jiacheng Liu, Saumya Malik, William Merrill, Lester James V. Miranda, Jacob Morrison, Tyler Murray, Crystal Nam, Valentina Pyatkin, Aman Rangapur, Michael Schmitz, Sam Skjonsberg, David Wadden, Christopher Wilhelm, Michael Wilson, Luke Zettlemoyer, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. 2025. 2 olmo 2 furious. *Preprint*, arXiv:2501.00656.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, Alex Renzin, Alex Tachard Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoochian, Amin Tootoonchian, Ananya Kumar, Andrea Vallone, Andrej Karpathy, Andrew Braunstein, Andrew Cann, Andrew Codispoti, Andrew Galu, Andrew Kondrich, Andrew Tulloch, Andrey Mishchenko, Angela Baek, Angela Jiang, Antoine Pelisse, Antonia Woodford, Anuj Gosalia, Arka Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver, Barret Zoph, Behrooz Ghorbani, Ben Leimberger, Ben Rossen, Ben Sokolowsky, Ben Wang, Benjamin Zweig, Beth Hoover, Blake Samic, Bob McGrew, Bobby Spero, Bogo Giertler, Bowen Cheng, Brad Lightcap, Brandon Walkin, Brendan Quinn, Brian Guarraci, Brian Hsu, Bright Kellogg, Brydon Eastman, Camillo Lugaresi, Carroll Wainwright, Cary Bassin, Cary Hudson, Casey Chu, Chad Nelson, Chak Li, Chan Jun Shern, Channing Conger, Charlotte Barette, Chelsea Voss, Chen Ding, Cheng Lu, Chong Zhang, Chris Beaumont, Chris Hallacy, Chris Koch, Christian Gibson, Christina Kim, Christine Choi, Christine McLeavey, Christopher Hesse, Claudia Fischer, Clemens Winter, Coley Czarnecki, Colin Jarvis, Colin Wei, Constantin Koumouzelis, Dane Sherburn, Daniel Kappler, Daniel Levin, Daniel Levy, David Carr, David Farhi, David Mely, David Robinson, David Sasaki, Denny Jin, Dev Valladares, Dimitris Tsipras, Doug Li, Duc Phong Nguyen, Duncan Findlay, Edede Oiwoh, Edmund Wong, Ehsan Asdar, Elizabeth Proehl, Elizabeth Yang, Eric Antonow, Eric Kramer, Eric Peterson, Eric Sigler, Eric Wallace, Eugene Brevdo, Evan Mays, Farzad Khorasani, Felipe Petroski Such, Filippo Raso, Francis Zhang, Fred von Lohmann, Freddie Sulit, Gabriel Goh, Gene Oden, Geoff Salmon, Giulio Starace, Greg Brockman, Hadi Salman, Haiming Bao, Haitang Hu, Hannah Wong, Haoyu Wang, Heather Schmidt, Heather Whitney, Heewoo Jun, Hendrik Kirchner, Henrique Ponde de Oliveira Pinto, Hongyu Ren, Huiwen Chang, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian O'Connell, Ian Osband, Ian Silber, Ian Sohl, Ibrahim Okuyucu, Ikai Lan, Ilya Kostrikov, Ilya Sutskever, Ingmar Kanitscheider, Ishaan Gulrajani, Jacob Coxon, Jacob Menick, Jakub Pachocki, James Aung, James Betker, James Crooks, James Lennon, Jamie Kiros, Jan Leike, Jane Park, Jason Kwon, Jason Phang, Jason Teplitz, Jason Wei, Jason Wolfe, Jay Chen, Jeff Harris, Jenia Varavva, Jessica Gan Lee, Jessica Shieh, Ji Lin, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joanne Jang, Joaquin Quinonero Candela, Joe Beutler, Joe Landers, Joel Parish, Johannes Heidecke, John Schulman, Jonathan Lachman, Jonathan McKay, Jonathan Uesato, Jonathan Ward, Jong Wook Kim, Joost Huizinga, Jordan Sitkin, Jos Kraaijeveld, Josh Gross, Josh Kaplan, Josh Snyder, Joshua Achiam, Joy Jiao, Joyce Lee, Juntang Zhuang, Justyn Harriman, Kai Fricke, Kai Hayashi, Karan Singhal, Katy Shi, Kavin Karthik, Kayla Wood, Kendra Rimbach, Kenny Hsu, Kenny Nguyen, Keren Gu-Lemberg, Kevin Button, Kevin Liu, Kiel Howe, Krithika Muthukumar, Kyle Luther, Lama Ahmad, Larry Kai, Lauren Itow, Lauren Workman, Leher Pathak, Leo Chen, Li Jing, Lia Guy, Liam Fedus, Liang Zhou, Lien Mamitsuka, Lilian Weng, Lindsay McCallum, Lindsey Held, Long Ouyang, Louis Feuvrier, Lu Zhang, Lukas Kondraciuk, Lukasz Kaiser, Luke Hewitt, Luke Metz, Lyric Doshi, Mada Aflak, Maddie Simens, Madelaine Boyd, Madeleine Thompson, Marat Dukhan, Mark Chen, Mark Gray, Mark Hudnall, Marvin Zhang, Marwan Aljubeh, Mateusz Litwin, Matthew Zeng, Max Johnson, Maya Shetty, Mayank Gupta, Meghan Shah, Mehmet Yatbaz, Meng Jia Yang, Mengchao Zhong, Mia Glaese, Mianna Chen, Michael Janner, Michael Lampe, Michael Petrov, Michael Wu, Michele Wang, Michelle Fradin, Michelle Pokrass, Miguel Castro, Miguel Oom Temudo de Castro, Mikhail Pavlov, Miles Brundage, Miles Wang, Minal Khan, Mira Murati, Mo Bavarian, Molly Lin, Murat Yesildal, Nacho Soto, Natalia Gimelshein, Natalie Cone, Natalie Staudacher, Natalie Summers, Natan LaFontaine, Neil Chowdhury, Nick Ryder, Nick Stathas, Nick Turley, Nik Tezak, Niko Felix, Nithanth Kudige, Nitish Keskar, Noah Deutsch, Noel Bundick, Nora Puckett, Ofir Nachum, Ola Okelola, Oleg Boiko, Oleg Murk, Oliver Jaffe, Olivia Watkins, Olivier Godement, Owen Campbell-Moore, Patrick Chao, Paul McMillan, Pavel Belov, Peng Su, Peter Bak, Peter Bakkum, Peter Deng, Peter Dolan, Peter Hoeschele, Peter Welinder, Phil Tillet, Philip Pronin, Philippe Tillet, Prafulla Dhariwal, Qiming Yuan, Rachel Dias, Rachel Lim, Rahul Arora, Rajan Troll, Randall Lin, Rapha Gontijo Lopes, Raul Puri, Reah Miyara, Reimar Leike, Renaud Gaubert, Reza Zamani, Ricky Wang, Rob Donnelly, Rob Honsby, Rocky Smith, Rohan Sahai, Rohit Ramchandani, Romain Huet, Rory Carmichael, Rowan Zellers, Roy Chen, Ruby Chen, Ruslan Nigmatullin, Ryan Cheu, Saachi Jain, Sam Altman, Sam Schoenholz, Sam Toizer, Samuel Miserendino, Sandhini Agarwal, Sara Culver, Scott Ethersmith, Scott Gray, Sean Grove, Sean Metzger, Shamez Hermani, Shantanu Jain, Shengjia Zhao, Sherwin Wu, Shino Jomoto, Shirong Wu, Shuaiqi, Xia, Sonia Phene, Spencer Papay, Srinivas Narayanan, Steve Coffey, Steve Lee, Stewart Hall, Suchir Balaji, Tal Broda, Tal Stramer, Tao Xu, Tarun Gogineni, Taya Christianson, Ted Sanders, Tejal Patwardhan, Thomas Cunninghman, Thomas Degry, Thomas Dimson, Thomas Raoux, Thomas Shadwell, Tianhao Zheng, Todd Underwood, Todor Markov, Toki Sherbakov, Tom Rubin, Tom Stasi, Tomer Kaftan, Tristan Heywood, Troy Peterson, Tyce Walters, Tyna Eloundou, Valerie Qi, Veit Moeller, Vinnie Monaco, Vishal Kuo, Vlad Fomenko, Wayne Chang, Weiyi Zheng, Wenda Zhou, Wesam Manassra, Will Sheu, Wojciech Zaremba, Yash Patil, Yilei Qian, Yongjik Kim, Youlong Cheng, Yu Zhang, Yuchen He, Yuchen Zhang, Yujia Jin, Yunxing Dai, and Yury Malkov. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

OpenAI. 2024. Learning to reason with llms.

Yonatan Oren, Nicole Meister, Niladri Chatterji, Faisal Ladhak, and Tatsunori B Hashimoto. 2023. Proving test set contamination in black box language models. *arXiv preprint arXiv:2310.17623*.

Kiho Park, Yo Joong Choe, and Victor Veitch. 2024. The linear representation hypothesis and the geometry of large language models. *Preprint*, arXiv:2311.03658.

Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, and Vedant Misra. 2022. Grokking: Generalization beyond overfitting on small algorithmic datasets. *arXiv preprint arXiv:2201.02177*.

Akshara Prabhakar, Thomas L. Griffiths, and R. Thomas McCoy. 2024. Deciphering the factors influencing the efficacy of chain-of-thought: Probability, memorization, and noisy reasoning. *Preprint*, arXiv:2407.01687.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. 2024. Steering llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15504–15522, Bangkok, Thailand. Association for Computational Linguistics.

Baptiste Roziere, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xiaoqing Ellen Tan, Yossi Adi, Jingyu Liu, Romain Sauvestre, Tal Remez, et al. 2023. Code llama: Open foundation models for code. *arXiv preprint arXiv:2308.12950*.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. 2022. Language models are multilingual chain-of-thought reasoners. *Preprint*, arXiv:2210.03057.

Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. 2024. Detecting pretraining data from large language models. In *The Twelfth International Conference on Learning Representations*.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2024. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36.

Asa Cooper Stickland, Alexander Lyzhov, Jacob Pfau, Salsabila Mahdi, and Samuel R Bowman. 2024. Steering without side effects: Improving post-deployment control of language models. *arXiv preprint arXiv:2406.15518*.

Alessandro Stolfo, Yonatan Belinkov, and Mrinmaya Sachan. 2023a. A mechanistic interpretation of arithmetic reasoning in language models using causal mediation analysis. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7035–7052.

Alessandro Stolfo, Yonatan Belinkov, and Mrinmaya Sachan. 2023b. A mechanistic interpretation of arithmetic reasoning in language models using causal mediation analysis. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7035–7052, Singapore. Association for Computational Linguistics.

Xiaojuan Tang, Zilong Zheng, Jiaqi Li, Fanxu Meng, Song-Chun Zhu, Yitao Liang, and Muhan Zhang. 2023. Large language models are in-context semantic reasoners rather than symbolic reasoners. *arXiv preprint arXiv:2305.14825*.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stanczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Olivier Bachem, Alanna Walton, Aliaksei Severyn, Alicia Parrish, Aliya Ahmad, Allen Hutchison, Alvin Abdagic, Amanda Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony Laforge, Antonia Paterson, Ben Bastian, Bilal Piot, Bo Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry, Chris Welty, Christopher A. Choquette-Choo, Danila Sinopalnikov, David Weinberger, Dimple Vijaykumar, Dominika Rogozińska, Dustin Herbison, Elisa Bandy, Emma Wang, Eric Noland, Erica Moreira, Evan Senter, Evgenii Eltyshev, Francesco Visin, Gabriel Rasskin, Gary Wei, Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna Klimczak-Plucińska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Peng Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost van Amersfoort, Josh Gordon, Josh Lipschultz, Josh Newlan, Ju yeong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjoesund, Lauren Usui, Laurent Sifre, Lena Heuermann, Leticia Lago, Lilly McNealus, Livio Baldini Soares, Logan Kilpatrick, Lucas Dixon, Luciano Martins, Machel Reid, Manvinder Singh, Mark Iverson, Martin Görner, Mat Velloso, Mateo Wirth, Matt Davidow, Matt Miller, Matthew Rahtz, Matthew Watson, Meg Risdal, Mehran Kazemi, Michael Moynihan, Ming Zhang, Minsuk Kahng, Minwoo Park, Mofi Rahman, Mohit Khatwani, Natalie Dao, Nenshad Bardoliwalla, Nesh Devanathan, Neta Dumai, Nilay Chauhan, Oscar Wahltinez, Pankil Botarda, Parker Barnes, Paul Barham, Paul Michel, Pengchong Jin, Petko Georgiev, Phil Culliton, Pradeep Kuppala, Ramona Comanescu, Ramona Merhej, Reena Jana, Reza Ardeshir Rokni, Rishabh Agarwal, Ryan

Mullins, Samaneh Saadat, Sara Mc Carthy, Sarah Cogan, Sarah Perrin, Sébastien M. R. Arnold, Sebastian Krause, Shengyang Dai, Shruti Garg, Shruti Sheth, Sue Ronstrom, Susan Chan, Timothy Jordan, Ting Yu, Tom Eccles, Tom Hennigan, Tomas Kocisky, Tulsee Doshi, Vihan Jain, Vikas Yadav, Vilobh Meshram, Vishal Dharmadhikari, Warren Barkley, Wei Wei, Wenming Ye, Woohyun Han, Woosuk Kwon, Xiang Xu, Zhe Shen, Zhitao Gong, Zichuan Wei, Victor Cotruta, Phoebe Kirk, Anand Rao, Minh Giang, Ludovic Peran, Tris Warkentin, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, D. Sculley, Jeanine Banks, Anca Dragan, Slav Petrov, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Sebastian Borgeaud, Noah Fiedel, Armand Joulin, Kathleen Kenealy, Robert Dadashi, and Alek Andreev. 2024. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

Curt Tigges, Oskar John Hollinsworth, Atticus Geiger, and Neel Nanda. 2023. Linear representations of sentiment in large language models. *arXiv preprint arXiv:2310.15154*.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. 2024. Language models don't always say what they think: unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Boshi Wang, Xiang Yue, Yu Su, and Huan Sun. 2024a. Grokked transformers are implicit reasoners: A mechanistic journey to the edge of generalization. *Preprint*, arXiv:2405.15071.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. 2024b. MMLU-pro: A more robust and challenging multi-task language understanding benchmark. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Zhaofeng Wu, Linlu Qiu, Alexis Ross, Ekin Akyürek, Boyuan Chen, Bailin Wang, Najoung Kim, Jacob Andreas, and Yoon Kim. 2024. Reasoning or reciting? exploring the capabilities and limitations of language models through counterfactual tasks. *Preprint*, arXiv:2307.02477.

Chulin Xie, Yangsibo Huang, Chiyuan Zhang, Da Yu, Xinyun Chen, Bill Yuchen Lin, Bo Li, Badih Ghazi, and Ravi Kumar. 2024. On memorization of large language models in logical reasoning. *Preprint*, arXiv:2410.23123.

Sohee Yang, Elena Gribovskaya, Nora Kassner, Mor Geva, and Sebastian Riedel. 2024. Do large language models latently perform multi-hop reasoning? *arXiv preprint arXiv:2402.16837*.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2024. Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*, 36.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*.

Tian Ye, Zicheng Xu, Yuanzhi Li, and Zeyuan Allen-Zhu. 2024. Physics of language models: Part 2.1, grade-school math and the hidden reasoning process. *Preprint*, arXiv:2407.20311.

Lei Yu, Virginie Do, Karen Hambardzumyan, and Nicola Cancedda. 2024. Robust llm safeguarding via refusal feature adversarial training. *arXiv preprint arXiv:2409.20089*.

Matej Zečević, Moritz Willig, Devendra Singh Dhami, and Kristian Kersting. 2023. Causal parrots: Large language models may talk causality but are not causal. *Preprint*, arXiv:2308.13067.

Chiyuan Zhang, Daphne Ippolito, Katherine Lee, Matthew Jagielski, Florian Tramèr, and Nicholas Carlini. 2023. Counterfactual memorization in neural language models. *Advances in Neural Information Processing Systems*, 36:39321–39362.

Yuze Zhao, Tianyun Ji, Wenjun Feng, Zhenya Huang, Qi Liu, Zhiding Liu, Yixiao Ma, Kai Zhang, and Enhong Chen. 2025. Unveiling the magic of code reasoning through hypothesis decomposition and amendment. In *The Thirteenth International Conference on Learning Representations*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

Zifan Zheng, Yezhaohui Wang, Yuxin Huang, Shichao Song, Mingchuan Yang, Bo Tang, Feiyu Xiong, and Zhiyu Li. 2024. Attention heads of large language models: A survey. *Preprint*, arXiv:2409.03752.

A Prompts

Table 3 presents the prompt we used to query GPT-4o to assign a Reasoning Score to each question.

B Details of Datasets

Here, we provide further details about the datasets used in Sections 3 and 4.

MMLU-Pro-M (Wang et al., 2024b) and MMLU-Pro-R MMLU-Pro is a comprehensive benchmark designed to assess the advanced language understanding and reasoning capabilities of large language models (LLMs). It spans 14 diverse domains such as mathematics, physics, chemistry, law, engineering, psychology, and health, encompassing over 12,000 questions. It features 10 options per question, significantly increasing the difficulty and robustness of the benchmark. Unlike MMLU, MMLU-Pro focuses on more challenging college-level problems that require deliberate reasoning across various domains. In this work, we use GPT-4o to assign a Reasoning Score to each question. We then divide the questions into two subsets: those with a score greater than 0.5 are categorized as MMLU-Pro-R, while those with a score of 0.5 or below are classified as MMLU-Pro-M.

PopQA (Mallen et al., 2023) PopQA focuses on evaluating factual knowledge in large language models, specifically targeting knowledge about entities, defined as triplets of (subject, relationship, object). The task is framed as open-domain question answering, where a model is asked to predict an answer without pre-given ground-truth paragraphs. This study explores few-shot learning and prompts LMs without parameter updates, in contrast to fine-tuning approaches. The performance is measured by accuracy, where a prediction is considered correct if any substring matches a gold answer.

C-Eval-H (Huang et al., 2023) C-EVAL is a comprehensive Chinese evaluation suite designed to assess the advanced knowledge and reasoning abilities of large language models (LLMs) in a Chinese context. As traditional NLP benchmarks primarily focus on English and fail to capture the unique challenges of Chinese language models, C-EVAL addresses this gap by providing a detailed evaluation framework tailored to the Chinese language and culture. It includes 13,948 multiple-choice questions across 52 diverse disciplines, ranging from humanities to science and engineering, and spans four difficulty levels: middle school, high school, college, and professional exams. In this work, we focus on the humanities portion and refer to it as C-Eval-H.

GSM8k (Cobbe et al., 2021) GSM8k is a dataset designed to evaluate the mathematical reasoning abilities of large language models (LLMs). It consists of 8.5K grade school-level math problems paired with natural language solutions. The dataset aims to address the challenges faced by LLMs in performing multi-step mathematical reasoning, which often reveals a critical weakness in these models.

MGSM (Shi et al., 2022) The MGSM (Multilingual Grade School Math) benchmark is introduced to assess multilingual reasoning abilities in large language models, addressing the gap between English-based chain-of-thought (COT) reasoning and multilingual NLP tasks. Building on the GSM8K dataset, MGSM extends it to ten typologically diverse languages through manual translations. .

C Additional Experiments

C.1 Detailed Plot of the PCA results

In this section, we present additional PCA results from various layers of the LLaMA3-8B-base and Gemma2-9B-base models discussed in Section 3.2, which is shown in Figure 7 and Figure 8. We also provide fine-grained PCA visualizations of questions from different subject domains in MMLU-Pro in Figure 9. Additionally, we include heatmaps in Figure 10 demonstrating that the first principal component from our PCA experiments captures the majority of the mean activation differences between $\mathcal{D}_{\textbf{Memory}}$ and $\mathcal{D}_{\textbf{Reasoning}}$.

D Details of the Intervention Experiments

Here, we provide more implementation details in the Features Intervention Experiments described in Section 4.

Inference Settings For the few-shot settings, we adhere to the original experimental setup across all datasets. Specifically, we use 5-shot for MMLU-Pro-M, MMLU-Pro-R, and C-Eval-H, and 8-shot for GSM8k, MGSM, and GSM-Symbolic. Additionally, we run 0-shot for PopQA, following the original configuration.

Prompt

- **Analyze the question to determine its position on the reasoning-memory spectrum. Return:**

  1. Concise justification (1-2 sentences)
  2. Score [0.0–1.0] where:
     - 1.0 = Strictly requires multi-step reasoning (calculations/formulas/deductions)
     - 0.0 = Purely factual recall or the inference of humanities knowledge
     - Intermediate values indicate hybrid characteristics

  Scoring Guidelines:

  - +0.5 if contains numerical values/percentages
  - +0.3 per required calculation step
  - +0.2 if requires unit conversions
  - -0.4 if answer appears verbatim in STEM textbooks
  - Max 1.0 | Min 0.0

  Examples:

  1. **Score 0.0:**
     Question: "Polarization is a property of..."
     Options: [transverse waves,...]
     Analysis: Directly tests textbook knowledge about wave properties without calculations.
     Score: 0.0
  2. **Score 0.35:**
     Question: "An owner of an apartment building in a rundown section of town knew...If the neighbor asserts a claim against the owner to recover damages for his injury, he should"
     Options: [not recover, because the owner can't be held responsible...]
     Analysis: Humanities-oriented question, which, although requiring multi-step reasoning, still leans more towards a memorization-based approach.
     Score: 0.35
  3. **Score 0.95:**
     Question: "Order from greatest to least: 3, 3 and 1 over 8, 3.8, 3.18."
     Options: ['3.8, 3 and 1 over 8, 3.18, 3',...]
     Analysis: Requires comparing numerical values and determining their order.
     Score: 0.95

  Current Analysis:

  Question: "{question_text}"
  Options: {options_list}
  Analysis:

Table 3: Prompt used to query GPT-4o to assign a Reasoning Score to each question.

For both open-ended generation and multi-choices question answering tasks, we allow the model to generate the next 200 tokens.

Validation-Test Set Split For parameter tuning and inference, we directly utilized the pre-existing validation and test sets that were already partitioned within each dataset.

Hyperparameters Selection Based on the validation and test sets we have split, we tune the hyperparameter, α, on the validation set. We adjust it in intervals of 0.05 in absolute value and select the value of α that performs best on the validation set to apply to the test set.

All the experiments in this work were conducted on four 80GB NVIDIA A800 GPUs.

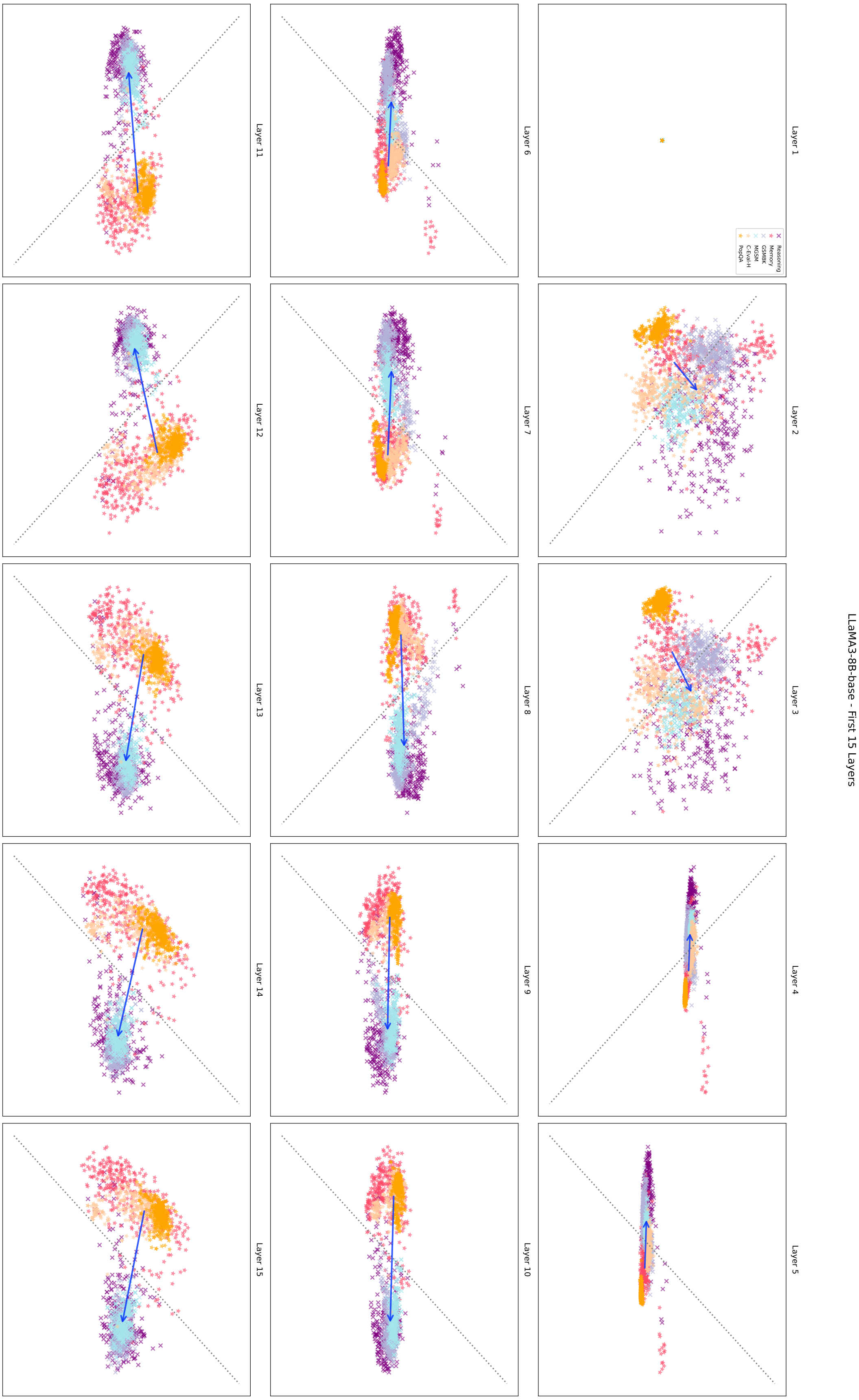

Figure 7: The PCA experiments results on the first 15 layers on LLaMA3-8B-base models

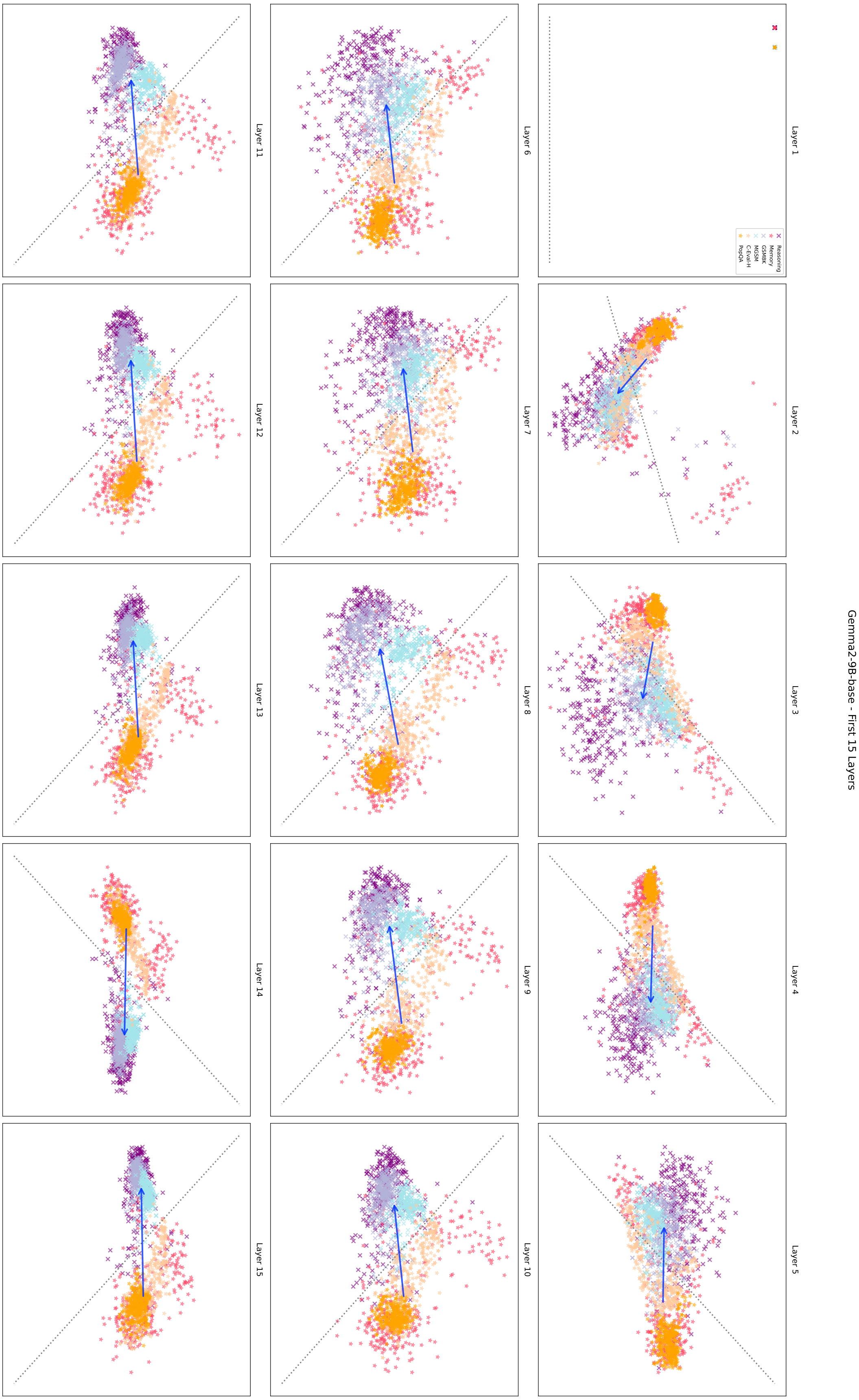

Figure 8: The PCA experiments results on the first 15 layers on Gemma2-9B-base models

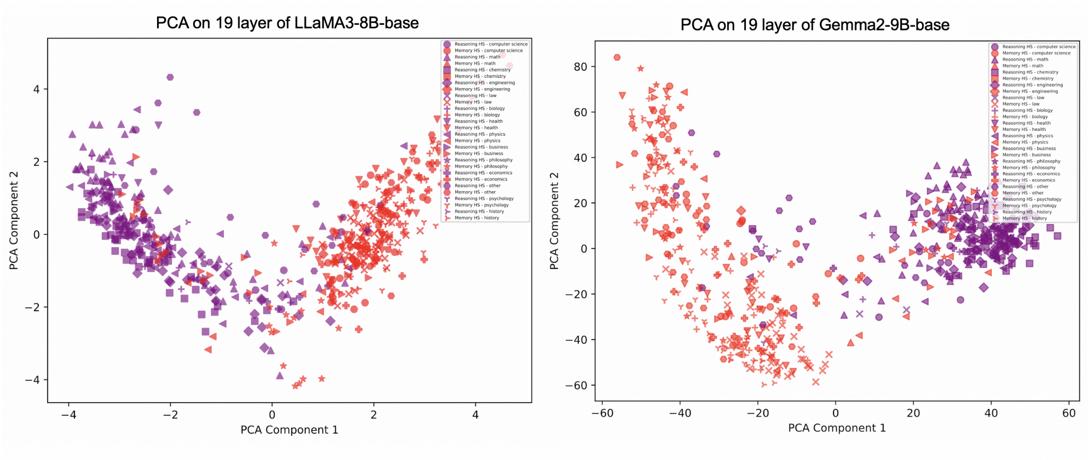

Figure 9: Fine-grained PCA visualizations of questions from different subject domains in MMLU-Pro on the model of LLaMA3-8B-base and Gemma2-9B-base.

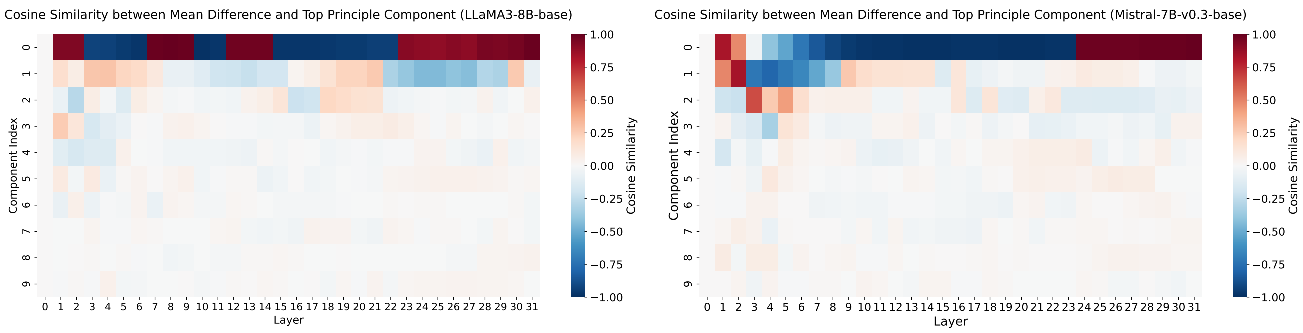

Figure 10: The top one principal component in PCA experiments captures most of the mean difference (Equation 2 between the activations in $\mathcal{D}_{\textbf{Memory}}$ and $\mathcal{D}_{\textbf{Reasoning}}$.