The Rise of Parameter Specialization for Knowledge Storage in Large Language Models

Yihuai Hong[1,2*] **Yiran Zhao**[3] **Wei Tang**[1] **Yang Deng**[4] **Yu Rong**[1] **Wenxuan Zhang**[5†]

[1]Alibaba DAMO Academy [2]New York University
[3]National University of Singapore
[4]Singapore Management University [5]Singapore University of Technology and Design

yihuaihong@nyu.edu, wxzhang@sutd.edu.sg

Abstract

Over time, a growing wave of large language models from various series has been introduced to the community. Researchers are striving to maximize the performance of language models with constrained parameter sizes. However, from a microscopic perspective, there has been limited research on how to better store knowledge in model parameters, particularly within MLPs, to enable more effective utilization of this knowledge by the model. In this work, we analyze twenty publicly available open-source large language models to investigate the relationship between their strong performance and the way knowledge is stored in their corresponding MLP parameters. Our findings reveal that as language models become more advanced and demonstrate stronger knowledge capabilities, their parameters exhibit increased specialization. Specifically, parameters in the MLPs tend to be more focused on encoding similar types of knowledge. We experimentally validate that this specialized distribution of knowledge contributes to improving the efficiency of knowledge utilization in these models. Furthermore, by conducting causal training experiments, we confirm that this specialized knowledge distribution plays a critical role in improving the model's efficiency in leveraging stored knowledge.

1 Introduction

An increasing number of powerful large language models (LLMs) have emerged in recent years (Touvron et al., 2023a; Achiam et al., 2023; Groeneveld et al., 2024; Bai et al., 2023; Team, 2025), often demonstrating remarkable capabilities across various benchmarks and tests (Hendrycks et al., 2021a; Chen et al., 2021a; Cobbe et al., 2021). Thanks to the large parameter space, they have shown an exceptional ability to encode vast amounts of knowledge within their parameters, enabling superior performance on knowledge-intensive tasks (Hendrycks et al., 2021a; Zhang et al., 2023).

To understand the internal mechanism of knowledge storage, many studies have been conducted. For example, Geva et al. (2021b) interprets the MLP layers of the transformer architecture (Vaswani et al., 2017) as key-value memories, where the factual knowledge encoded in the weights is retrieved and transmitted to the output layer during inference (Geva et al., 2023; Meng et al., 2022; Yu et al., 2024). Furthermore, researchers have observed that, in the final layer of the MLP, each vector in that value matrix can act as a fundamental unit of knowledge storage (Geva et al., 2022a,b). However, there has been limited research on how to better store and compress knowledge within constrained model parameters to enable more effective utilization of that knowledge by the model.

[*]Work done during an internship at Alibaba Group, before joining New York University.
[†]Corresponding author.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

In this work, we investigate the relationship between language models' knowledge storage patterns and their performance. To identify parameters associated with specific knowledge concepts, we analyze consistently activated parameters in MLP layers when the model processes questions related to the particular knowledge concept. Building on the key-value interpretation of the MLP by Geva et al. (2021b), which treats the up-projection matrix as the key and the down-projection matrix as the value (*i.e.*, stored knowledge), we extract the intermediate representations between these two matrices and treat their absolute value as the activation of corresponding parameters. To support empirical analysis, we construct a new encyclopedic knowledge benchmark based on Wikipedia, covering knowledge concepts with varying frequencies. We then apply the knowledge parameter identification method to 20 open-source LLMs across a wide range of model families, enabling us to explore correlations between knowledge storage patterns and overall model performance.

Our extensive empirical analysis reveals that **stronger models exhibit higher parameter specialization for distinct knowledge**, whereas weaker models distribute knowledge more diffusely across parameters. Consequently, significantly more parameters are required to store individual knowledge in weaker models. As illustrated in Figure 1, advancing model capability correlated with improved parameter specialization for encoding knowledge: fewer parameters are allocated per knowledge concept, while each parameter governs a narrower subset of concepts.

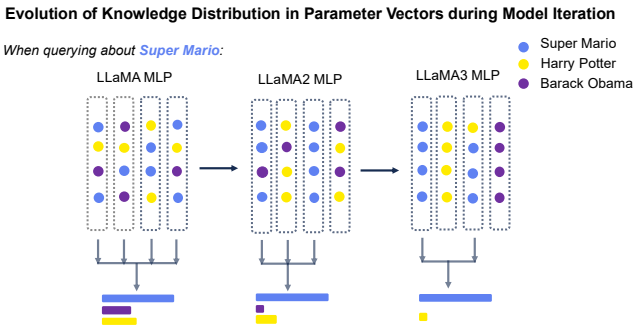

Figure 1: Evolution of knowledge distribution in model parameters during three iterations of LLaMA models. Each parameter vector corresponds to a column in the value matrix of the MLP module, as indicated by the dashed rectangles.

Motivated by this observation, we further conduct four sets of controlled experiments, each involving continued training on the Llama2-7B (Touvron et al., 2023a) and Qwen2-7B (Yang et al., 2024) models with new knowledge respectively, to validate the strong causal relationship between improved parameter specialization and enhanced performance of the models on knowledge tasks. Overall, the experiments reveal that encoding similar knowledge into the same parameter vectors better aligns with the model's internal knowledge retrieval mechanism. This approach helps the model utilize knowledge more efficiently, improves knowledge compression, and reduces hallucination generation.

Our contributions can be summarized as follows:

- To the best of our knowledge, this is the first attempt to quantify and compare the degree of parameter specialization for knowledge storage across different LLMs.
- We investigate the relationship between parameter specialization and model performance in LLMs, constructing a dedicated probing dataset for an in-depth analysis on 20 open-source LLMs. Our findings indicate that more capable LLMs exhibit greater parameter specialization.
- Through controlled training experiments, we provide empirical evidence of a causal link between increased parameter specialization and improved performance on knowledge-intensive tasks.

2 Related Work

Knowledge Storage in LLMs Studying how knowledge is stored and utilized in LLMs has been an important area in the research of LLM interpretability (Meng et al., 2022; Geva et al., 2021b; Sukhbaatar et al., 2015; Geva et al., 2023). Recent studies have shown that MLPs are the primary and crucial components for storing factual knowledge and associations in transformer-based language models (Geva et al., 2022b; Dar et al., 2023). They can be conceptualized as key-value memories (Geva et al., 2021b), where the factual knowledge encoded in the MLP weights is recalled and transmitted to the output layer during inference (Geva et al., 2023; Meng et al., 2022; Yu et al., 2024). Additionally, researchers have found that in the final layer of the MLP, each vector in the value matrix can serve as a fundamental unit for storing knowledge (Geva et al., 2022a,b). They have also verified

that by directly manipulating or disrupting these parameter vectors, specific knowledge can be edited or unlearned (Hong et al., 2024a,b; Meng et al., 2022), leading to changes in the model's responses.

Knowledge Superposition in LLM Elhage et al. (2022); Olah (2023) propose the concept of Knowledge Superposition. It refers to an inevitable phenomenon in neural network models, especially large language models, during training and data memorization: since the number of data features greatly exceeds the number of parameters in the model, each parameter does not have a simple one-to-one mapping with the data features or knowledge. Neurons are often involved with multiple data features simultaneously. In our work, we treat each vector in the last layer of MLP as a basic unit for storing knowledge and investigate the superposition of knowledge within these vectors.

3 Parameter Specialization Analysis for Knowledge Storage

3.1 Preliminary

In transformer-based language models, the MLP is a crucial component for storing the model's factual knowledge, and its sub-layers can be viewed as key-value memories (Geva et al., 2021b). To be specific, the first layer* of MLP sublayers can be viewed as a matrix W_K formed by key vectors $\{\mathbf{k}_1, \mathbf{k}_2, \ldots, \mathbf{k}_n\}$, used to capture a set of patterns in the input sequence, and ultimately outputting the coefficient scores. The second layer can be viewed as a matrix W_V formed by value vectors $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}$, with each value vector containing the corresponding factual knowledge.

Formally, the output of the MLP in the transformer's ℓ-th layer, given an input hidden state $\mathbf{x}^\ell$, can be defined as:

$$\mathbf{M}^\ell = f\big(W_K^\ell \cdot \gamma(\mathbf{x}^\ell + \mathbf{A}^\ell)\big) W_V^\ell = \mathbf{m}^\ell W_V^\ell, \tag{1}$$

where $W_K^\ell, W_V^\ell \in \mathbb{R}^{n \times d}$. The function f and γ represent a non-linearity† and layer normalization, respectively. In the transformer's ℓ-th layer, $\mathbf{m}^\ell \in \mathbb{R}^n$ denotes the coefficient scores, and $\mathbf{A}^\ell$ represents the output of the attention component. The hidden state dimension is d, while the intermediate MLP has a dimension of n. Then, by denoting $\mathbf{v}_j^\ell$ as the j-th column (which will be called the value vector or parameter vector in the following sections) of W_V^ℓ and m_j^ℓ as the j-th element in the coefficients produced by the first layer of the MLP, we can view MLP's output $\mathbf{M}^\ell$ as a linear combination of the value vectors in W_V^ℓ, with their corresponding coefficients $\mathbf{m}^\ell$:

$$\mathbf{M}^\ell = \sum\nolimits_{j=1}^{n} m_j^\ell \mathbf{v}_j^\ell, \tag{2}$$

Finally, the hidden states at the ℓ-th layer of the language model can be defined as:

$$X^{\ell+1} = X^\ell + \mathbf{M}^\ell + \mathbf{A}^\ell, \tag{3}$$

where X^ℓ, $\mathbf{M}^\ell$ and $\mathbf{A}^\ell$ represent the hidden states, MLP's output, and the attention component's output in the transformer's ℓ-th layer, respectively. In this work, we focus on studying the impact of the MLP on the knowledge output of the hidden states.

3.2 Knowledge Vectors Masking Procedure

Referring to Eq. (2), if we aim to ablate the impact of the knowledge contained in the vectors for a particular subset S^ℓ of indices in ℓ-th layer, we can directly set the corresponding m_j^ℓ values for $j \in S^\ell$ to zero. Hence, we have:

$$\mathbf{M}_{\text{masked}}^\ell = \sum\nolimits_{\substack{j=1 \\ j \notin S^\ell}}^{n} m_j^\ell \mathbf{v}_j^\ell + \sum\nolimits_{\substack{j=1 \\ j \in S^\ell}}^{n} 0 \cdot \mathbf{v}_j^\ell = \sum\nolimits_{\substack{j=1 \\ j \notin S^\ell}}^{n} m_j^\ell \mathbf{v}_j^\ell, \tag{4}$$

*In most decoder-only models, such as GPT-2 (Radford et al., 2019) and GPT-J (Chen et al., 2021b), the MLP component consists of two layers, whereas in LLaMA (Touvron et al., 2023b), it comprises three layers. However, we can still regard LLaMA's first two layers collectively as the key matrices, with their output representing the coefficient scores.

†For brevity, the bias term is omitted.

Therefore, given a concept, when we aim to identify which specific value vectors in the model's MLPs are most closely related to the knowledge contained in that concept—while avoiding the masking of vectors associated with the model's general capabilities[‡] (Meng et al., 2022; Geva et al., 2023), i.e., determining the appropriate subset S^ℓ at each layer of the model for this concept—we will run t concept-related questions and t^* irrelevant questions on the selected model. Then we will compute the corresponding coefficients $\mathbf{m}^\ell$ and $\mathbf{m}^{*\ell}$, which are the averages of the coefficients for the concept-related questions and irrelevant questions, respectively, at each layer of the model. For details on the generation of concept-related and irrelevant questions, as well as the selection of t and t^*, please refer to §3.4. After obtaining $\mathbf{m}^\ell$ and $\mathbf{m}^{*\ell}$ at each layer, we perform the computation using the following formula:

$$\mathbf{S}^\ell = \left\{ \left| m_j^\ell - m_j^{*\ell} \right| \;\middle|\; 1 \le j \le n,\ m_j^\ell \in \mathbf{m}^\ell,\ m_j^{*\ell} \in \mathbf{m}^{*\ell} \right\} \tag{5}$$

Next, we will sort $\mathbf{S}^\ell$ in descending order and select the value vectors corresponding to the indices of the top k elements, which will be used as the subset S^ℓ for the masking operation. This allows us to observe and analyze the impact of masking these vectors on the model's knowledge output for certain concepts.

3.3 The Definition of Parameter Specialization

After obtaining the subset of value vectors S^ℓ that exhibit specificity to a given concept at each layer of the model, as described in §3.2, we apply the masking operation to these value vectors, as shown in Eq. (4). We then analyze its impact on the model's final outputs for the t concept-related questions and t^* irrelevant questions. By comparing the model's responses after masking with the ground truth answers, we compute the accuracy on concept-related questions, referred to as the Concept Specific Score after surgery, and the accuracy on irrelevant questions, referred to as the General Score after surgery. To quantify the degrees of specialization of the model's value vectors with respect to the concept-related knowledge, we define the **Parameter Specialization Score (PSS)**:

$$\text{PSS} \triangleq \frac{|\text{General Score after surgery} - \text{Concept Specific Score after surgery}|}{\text{General Score before surgery}}, \tag{6}$$

which is obtained by taking the absolute difference between the General Score and the Concept Specific Score after surgery, and then dividing by the model's accuracy on the entire dataset before surgery. A higher PSS indicates that the parameter vectors in the model's MLP layers exhibit a higher degree of specialization towards specific knowledge. Conversely, a lower PSS suggests more severe knowledge superposition phenomena within the parameter vectors, resulting in a lower degree of specialization.

3.4 Dataset Construction

To thoroughly investigate the parameter specialization of knowledge with different frequencies in the parameter vectors of LLMs' MLP, we introduce a dataset named SpecWiki. It includes 525 concepts selected from Wikipedia [§], a widely recognized high-quality corpus for LLM training. These concepts are categorized based on their frequency levels to ensure a diverse distribution. We then design two distinct question formats—multiple-choice questions and open-ended generation prompts—to facilitate a thorough examination of the models' knowledge storage.

Concept Selection We treat each Wikipedia item as a defining concept, typically represented by an article focused on a specific subject, indicated by its title. We focus on specific entity concepts, such as historical figures, events and locations. We began by randomly sampling 2,400 pages (a 0.01% rate) from the 2019 version of Wikipedia. Subsequently, we performed manual filtering to remove overly commonsensical or abstract concepts (such as the letter 'S' and the word 'Freedom'), ambiguous concepts (like 'Apple'), and those associated with pages under 1,000 words. Ultimately, this resulted in 525 high-quality concepts spanning specific topics like people, arts, and events.

[‡]The term "general ability" refers to the model's fundamental skills, such as processing text inputs correctly and generating coherent outputs, rather than encoding knowledge specific to a particular concept.

[§]https://en.wikipedia.org/

Given that the frequency of knowledge in training datasets significantly influences a model's ability to retain and comprehend it (Allen-Zhu & Li, 2023; Meng et al., 2022; Mallen et al., 2023), we utilize Wikipedia page views as a proxy for knowledge frequency in the models' pre-training datasets[¶]. To this end, we calculated the page views for each concept on Wikipedia between January 1, 2010, and December 31, 2019[‖]. Based on these statistics, concepts are categorized by page view frequency into three equal tiers: low-frequency (bottom 33% of the distribution), medium-frequency (middle 33%), and high-frequency (top 33%). A more detailed distribution of the categories and the corresponding example data of SpecWiki dataset are provided in Table 4 and Table 5, respectively, in the Appendix. This approximation helps estimate the likelihood of a concept's presence in the models' pre-training datasets and allows us to explore how knowledge at different frequency levels is stored in models and provides a more comprehensive evaluation.

Question Generation To more precisely assess the retention of knowledge within the model, we design two sets of question formats.

- *Multi-Choice Questions*. Drawing inspiration from the widely used Massive Multitask Language Understanding (MMLU) dataset (Hendrycks et al., 2021b), which evaluates general knowledge across models, we similarly designed ten multiple-choice questions for each concept, ensuring the knowledge and answers could be directly found in the relevant Wikipedia articles. Specifically, we provided GPT-4o (OpenAI et al., 2024) with the appropriate Wikipedia article for each concept and instructed it to extract ten questions without overlap, along with the correct answers derived from the article's text. Next, it was instructed to generate three additional incorrect answers, aside from the golden answer, ensuring that none of them overlapped with the correct answer to avoid confusion. The detailed prompt is available in §A.1. We also include sample multiple-choice questions and results of manual verification of the generated data in Appendix §A.2.
- *Open-ended Generation*. To more effectively assess the model's ability to generate knowledge text freely, and to overcome the randomness and lack of depth inherent in the Multi-Choice Question evaluation method, we also set up a series of Open-ended Generation questions. For each question related to a concept, we prompted the model to generate an answer of up to 150 tokens directly and used GPT-4o as an evaluator to evaluate whether the generated response correctly matched the golden answer.

4 Experiment

4.1 Experimental Setup

Evaluated Models To provide a more comprehensive evaluation of how the degree of parameter specialization evolves across large language models, we assessed 20 open-source models from various families and sizes in the community. Specifically, we evaluated LLaMA series (Touvron et al., 2023a,b; Grattafiori et al., 2024), Qwen series (Bai et al., 2023; Yang et al., 2024; Team, 2025), Gemma series (Team et al., 2024a,b), OLMo series (Groeneveld et al., 2024; OLMo et al., 2025), Yi series (AI et al., 2025), Mistral series (Jiang et al., 2023), GPT-j-6b (Wang & Komatsuzaki, 2021), Pythia-6.9b (Biderman et al., 2023), Falcon-7b (Almazrouei et al., 2023) and Mpt-7b (Databricks, 2023). Refer to Appendix §B.1 for the implementation details of these models.

Knowledge Vectors Masking Setup Based on the descriptions in §3.2, in order to obtain $\mathbf{m}^{\ell}$ and $\mathbf{m}^{*\ell}$ for each concept in SpecWiki at each layer of the model, we set the number of concept-related questions t to 10. Additionally, we randomly select 5 irrelevant concepts with no knowledge overlap from the benchmark, and gather the corresponding questions associated with these irrelevant concepts, resulting in $t^* = 50$ irrelevant questions. These collected questions will also be directly utilized in the computation of both the Concept-Specific Score and the General Score.

[¶]To better support this point, we include experiments in §A.4 of Appendix that validate the strong correlation between concept popularity and their frequency in the pretraining data.

[‖]The earliest release date of the evaluated models, such as GPT-J, is 2019. Therefore, their pretraining datasets could not include knowledge or concepts that emerged after this period. To ensure a fair evaluation across all models, any knowledge introduced post-2019, including the COVID-19 pandemic, was excluded from the benchmark.

Figure 2: Correlation between the performance on SpecWiki and parameter specialization score (PSS) in 20 language models. We use a color gradient to distinguish the release times of the models, with cooler colors indicating earlier release dates and warmer colors representing later releases. Additionally, the size of each circle reflects the model's performance on MMLU, with larger circles indicating better performance. The blue trendline, obtained through linear regression fitting of the data points, suggests a strong correlation between a model's performance on SpecWiki and its degree of Parameter Specialization.

Regarding the selection of model layers for masking, since the initial layers of a model typically handle fundamental capabilities like basic text processing (Meng et al., 2022; Geva et al., 2023), masking these layers could severely impair the model's basic text generation abilities. Therefore, for all models in our study, we preserve the first 5 layers without masking and only apply vector masking operations to all subsequent layers.

For the PSS computation of each model, we selected five different fixed k values—10%, 20%, 30%, 40%, and 50%—which represent the proportion of value vectors in the model's MLP layers that were masked. For each k, we calculated the corresponding PSS following 6, and then averaged the results to obtain the final PSS score for each model. This criterion was applied consistently across both the Multiple-Choice Questions (MCQ) and Open-ended Generation (OEG) tasks.

4.2 Main Results

The main results for the Multiple-Choice Questions setting can be seen in Figure 2. We observe a strong correlation between the degree of Parameter Specialization (measured by PSS) and model performance on SpecWiki across 20 models, with Pearson and Spearman coefficients of 0.92 and 0.93, respectively. Models achieving better performance on SpecWiki exhibit higher Parameter Specialization Scores. Furthermore, models with higher PSS are often those released more recently (warmer color) and exhibit stronger general abilities, as measured by their MMLU performance (larger circle). The corresponding results for the Open-ended Generation setting can be found in Figure 6 in §B.2, which exhibit similar patterns and trends.

To better analyze the variations in Parameter Specialization across models within the same family, we selected eight models from four model families: LLaMA, Qwen, Mistral, and Gemma. We examined how the difference between the General Score, which represents the model's ability to handle irrelevant knowledge, and the Concept Specific Score, which reflects the model's ability to handle task-specific knowledge, changes under different masking ratios of parameter vectors. The results are shown in Figure 3.

From the figure, we can observe a very similar pattern across models from the four families:

1. Among models within the same family, more advanced models tend to achieve higher peaks in the General Score - Concept Specific Score difference. This indicates that more advanced models generally exhibit higher levels of Parameter Specialization.

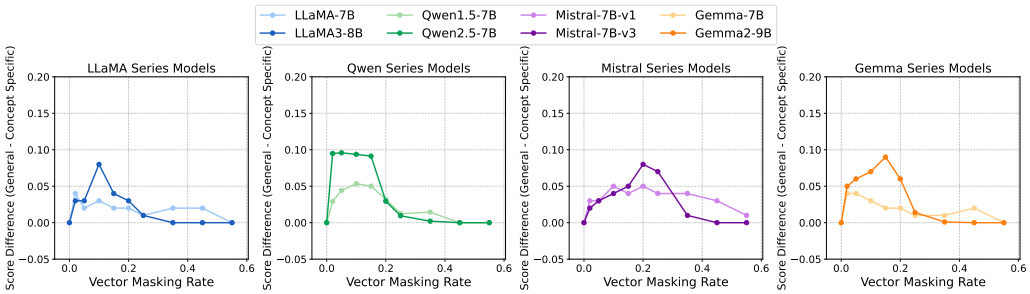

Figure 3: Analysis of Parameter Specialization variations across models within the same family. We selected eight models from four model families: LLaMA, Qwen, Mistral, and Gemma. The figure shows how the difference between the General Score (representing the model's ability to handle irrelevant knowledge) and the Concept Specific Score (representing the model's ability to handle task-specific knowledge) changes under different masking ratios of parameter vectors.

2. As the masking ratio of parameter vectors increases, from approximately 5% to 20%, the difference between the General Score and the Concept Specific Score gradually increases to a peak. This indicates that we are removing parameter vectors that are highly specific to the target knowledge. After reaching the peak, as the masking ratio continues to increase, the difference gradually decreases to zero. This suggests that parameter vectors with lower activation are often those that have a higher degree of knowledge superposition and are less specialized in the target knowledge.

Additionally, we unexpectedly found that when a small proportion of concept-related vectors (ranging from 5% to 10%) were masked, the performance of the masked models on unrelated questions even surpassed that of the original models. This observation is consistent across various models and indicates the positive impact of reducing irrelevant information interference in the model's representation, leading to improved performance.

In §5, we will further validate the causal relationship between the degree of model parameter specialization and its ability to better utilize target knowledge through the finetuning experiments on additional data.

4.3 Impact of Model Scale on Parameter Specialization

In this section, to better explore the differences in the degree of parameter specialization across models of different sizes, we conducted Knowledge Vectors Masking experiments on five Qwen1.5 models of varying sizes (0.5B, 1.8B, 4B, 7B, and 14B) and on 2 Gemma2 models (2B and 9B). The results are shown in Table 1. We observe that in both the Qwen and Gemma model families, as the model size increases, the corresponding Parameter Specialization Score also increases. This trend is accompanied by improved performance on SpecWiki. This suggests that in larger-scale models, the degree of superposition for specific knowledge decreases and it tends to be distinctly represented across designated parameter vectors.

Model	$\text{Accuracy}_{\text{MCQ}} \uparrow$	PSS ↑
Qwen1.5-0.5B	0.61 ($\pm$0.2)	0.019 ($\pm$0.01)
Qwen1.5-1.8B	0.61 ($\pm$0.3)	0.044 ($\pm$0.02)
Qwen1.5-4B	0.73 ($\pm$0.2)	0.075 ($\pm$0.02)
Qwen1.5-7B	0.75 ($\pm$0.2)	0.121 ($\pm$0.04)
Qwen1.5-14B	**0.82** ($\pm$0.2)	**0.184** ($\pm$0.03)
Gemma2-2B	0.72 ($\pm$0.3)	0.057 ($\pm$0.02)
Gemma2-9B	**0.86** ($\pm$0.1)	0.138 ($\pm$0.03)

Table 1: Performance comparison of language models with varying sizes on Multiple-Choice Question and Parameter Specialization Score. Both the Qwen1.5 and Gemma2 series models show improved Parameter Specialization as the model scale increases, accompanied by better performance on the MCQ testing in SpceWiki.

4.4 Evolution of Parameter Specialization During Pretraining

To better investigate the development of Parameter Specialization in the model from the perspective of model training dynamics, we analyzed 10 checkpoints from the OLMo-2-1124-7B (OLMo et al., 2025) pretraining process by using our SpecWiki. The results are shown in Figure 4 below.

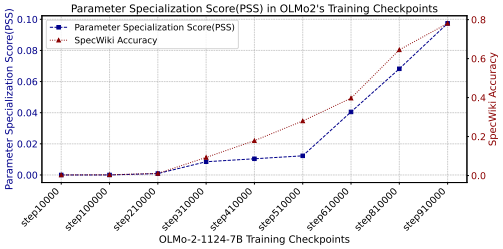

Figure 4: Development of Parameter Specialization in OLMo-2-1124-7B over the pretraining process.

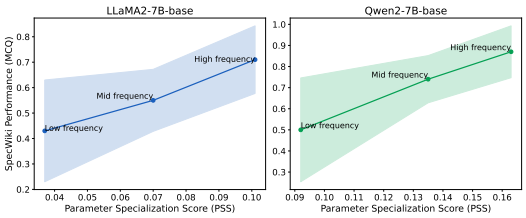

Figure 5: Relationship between concept popularity, model accuracy on MCQ, and Parameter Specialization Score in LLaMA2-7B and Qwen2-7B models.

Model	$\text{Accuracy}_{\text{MCQ}}$ ↑	$\text{Accuracy}_{\text{OEG}}$ ↑	PSS ↑	Semantic Entropy ↓	Local Intrinsic Dimension ↓
LLaMA2-7B	0.60 (±0.2)	0.51 (±0.1)	0.67 (±0.1)	0.67 (±0.1)	11.23 (±2.1)
LLaMA2-7B$_{FT-FV}$	0.63 (±0.3)	0.54 (±0.2)	0.65 (±0.2)	0.62 (±0.1)	11.12 (±1.4)
LLaMA2-7B$_{FT-PV}$	**0.67** (±0.3)	**0.59** (±0.2)	**0.72** (±0.1)	**0.50** (±0.2)	**7.89** (±2.1)
LLaMA2-7B$_{FT-CV}$	0.62 (±0.1)	0.51 (±0.1)	0.63 (±0.2)	0.62 (±0.1)	11.12 (±1.4)
LLaMA2-7B$_{FT-RV}$	0.58 (±0.2)	0.49 (±0.2)	0.65 (±0.2)	0.65 (±0.2)	11.07(±2.7)
Qwen2-7B	0.72 (±0.3)	0.63 (±0.1)	0.124 (±0.03)	0.56 (±0.1)	9.78 (±1.9)
Qwen2-7B$_{FT-FV}$	0.73 (±0.2)	0.67 (±0.2)	0.110 (±0.02)	0.59 (±0.2)	8.53 (±1.1)
Qwen2-7B$_{FT-PV}$	**0.77** (±0.1)	**0.70** (±0.1)	**0.133** (±0.03)	**0.39** (±0.1)	**6.92** (±1.3)
Qwen2-7B$_{FT-CV}$	0.73 (±0.2)	0.65 (±0.2)	0.114 (±0.03)	0.55 (±0.2)	8.78 (±1.6)
Qwen2-7B$_{FT-RV}$	0.71 (±0.2)	0.63 (±0.1)	0.122 (±0.02)	0.59 (±0.2)	9.65 (±1.4)

Table 2: The performance of both the original LLaMA2-7B-base and Qwen2-7B-base models, along with their FT-FV, FT-PV, FT-CV and FT-RV variants, was assessed on a selection of 10 high-frequency concepts from SpecWiki. Five metrics were used to evaluate their performance, including their general effectiveness, Parameter Specialization, and the degree of hallucination present in their output.

From the results, we observe that during the early training steps (step 10,000 to step 210,000), both the PSS and the accuracy on SpeciWiki remain nearly unchanged and close to zero. In the subsequent phase (step 310,000 to step 510,000), although the model begins to show noticeable gains in accuracy on SpeciWiki, the PSS are still under 0.1. However, it is during the later training steps (step 610,000 to step 910,000) that parameter specialization begins to emerge, accompanied by a more substantial improvement in accuracy. These findings suggest that parameter specialization does not occur in the early stages of training, but rather emerges after a certain amount of data exposure. Furthermore, as training continues and the model sees more data, the degree of parameter specialization increases accordingly.

4.5 Parameter Specialization in Relation to Concept Popularity

In this section, we analyze how the popularity of concepts themselves, which is roughly equivalent to their frequency in the pretraining data, will affect the level of parameter specialization for the corresponding knowledge. We follow the classification method for concepts as outlined in §3.4, dividing them into high-frequency, mid-frequency, and low-frequency categories. The impact on their PSS scores is measured on two example models, LLaMA2-7B, and Qwen2.5-7B, which are shown in Figure 5.

From the figure, it is clear that in both the LLaMA2-7B and Qwen2-7B models, as the popularity of a concept decreases, the model's accuracy on that specific knowledge declines, accompanied by a lower Parameter Specialization Score. This suggests that the degree of Parameter Specialization for a particular knowledge in the model's parameters is likely directly correlated with the frequency of that knowledge in the model's pretraining dataset. The higher the frequency, the greater the Parameter Specialization for that knowledge in the model.

5 Validation of Parameter Specialization Benefits for Knowledge Tasks

In this section, we conducted four sets of controlled training experiments, each involving continued fine-tuning on the Llama2-7B-base (Touvron et al., 2023a) and Qwen2-7B-base (Yang et al., 2024)

models with additional knowledge data. These experiments aim to validate the causal relationship between improved parameter specialization and enhanced model performance on knowledge tasks.

5.1 Finetuning Setup

We randomly selected 10 high-frequency concepts from the SpecWiki benchmark. For each concept, we gathered relevant textual material from the top 10 most popular Google search results, including the corresponding Wikipedia article, and compiled this into an additional finetuning training dataset.

Next, we will validate whether the improvement in Parameter Specialization and the enhanced efficiency in the model's use of knowledge truly exhibit a causal relationship through four distinct finetuning experiments. The experimental setups are detailed below:

FT-FV(Full Vectors) Perform full finetuning (FT) on all parameter vectors of the MLPs across all layers in the model, while keeping the other parameters frozen.

FT-PV(Partial Vectors) Perform partial finetuning on a subset of the parameter vectors in the MLPs while keeping the other parameters frozen. Specifically, for each model, we apply finetuning (FT) to the top $\frac{k}{8}$ most highly activated parameter vectors[**]. For the selection of k, please refer to the description in §4.1.

FT-CV(Complementary Vectors) Perform finetuning only on the complementary set of parameter vectors, excluding the target vectors.

FT-RV(Random Vectors) Perform finetuning on a subset of parameter vectors randomly selected from the MLP, ensuring the same quantity as in the FT-PV setting.

5.2 Finetuning Results

In addition to evaluating the model's performance on SpecWiki's Multi-choice Question and Open-ended Generation tests, as well as the Parameter Specialization scores, we also report two other metrics, Semantic Entropy (Kuhn et al., 2022) and Local Intrinsic Dimension (LID) (Yin et al., 2024), for measuring the extent of hallucination in the model's output. These metrics help evaluate whether training strategies that enhance Parameter Specialization—by aligning better with the model's knowledge retrieval mechanisms through a data-encoded strategy—can effectively reduce the unintended side effect of hallucination. For a detailed introduction to these two hallucination measurements, please refer to §B.3.

The final results are presented in Table 2. From the results, we can see that the FT-PV method, which finetunes only a small subset of the highly activated knowledge parameters, not only further enhances the model's Parameter Specialization compared to the three other finetuning setups, but also greatly improves the model's utilization of specific knowledge. As a result, it achieves the best performance on the benchmark Multi-Choice questions and Open-ended generation tasks. Additionally, by reducing the influence of irrelevant information in the model's key parameter vectors, FT-PV helps to significantly reduce the level of hallucination in the generated text.

Although FT-FV and FT-CV does improve the model's performance on both the Multi-Choice questions and Open-ended generation tasks to some extent, compared to FT-PV, it does not lead to a better increase in Parameter Specialization. Additionally, the degree of hallucination in the generated text is not effectively reduced. FT-RV, serving as a counterpart to FT-PV, demonstrates that fine-tuning the same number of arbitrary value vectors in the model's MLP can not result in a desirable knowledge enhancement.

6 Conclusion

This study reveals that enhanced parameter specialization—where related knowledge is encoded in focused parameter vectors—correlates with superior performance in large language models. Analyzing 20 open-source models, we observed stronger models increasingly consolidate similar knowledge into fewer parameters, while weaker models distribute it diffusely. Controlled experiments confirmed

[**]We experimented with $\frac{k}{2}$, $\frac{k}{4}$, $\frac{k}{8}$, and $\frac{k}{16}$, and found that finetuning only $\frac{k}{8}$ of the parameter vectors was sufficient to achieve excellent performance.

that optimizing this specialization improves task performance and reduces hallucination. These findings highlight the importance of aligning knowledge storage with models' retrieval mechanisms for efficiency and accuracy. Future work should explore dynamic knowledge updates and scalability, advancing both interpretability and performance in LLM design.

7 Limitations and Future Work

In our work, we have only examined and validated knowledge parameter specialization within the MLP, and this was done by treating vectors in the MLP as units of analysis. However, at least this remains one of the knowledge storage methods that has been extensively validated so far (Geva et al., 2021a; Meng et al., 2022; Geva et al., 2023). In fact, knowledge may also reside within the attention module of transformer models (Geva et al., 2023).

Additionally, due to GPU limitations, all the models we tested are smaller than or equal to 14B parameters, so we were unable to validate our conclusions on larger models, such as those with 35B parameters or more.

In future work, we will progressively narrow the focus of our research to individual neurons in language models, aiming to measure and validate more precise Parameter Specialization and Parameter Superposition. In addition, we will extend this concept to the study of other related model architectures, including Mixture of Experts (Fedus et al., 2022), which similarly enhances model performance by specializing expert parameters, as well as Sparse Auto-Encoders (Huben et al., 2024), which help clarify the model's representations by leveraging a larger parameter space and mitigating the superposition of these features.

Acknowledgment

This research is supported by the Ministry of Education, Singapore, under its Academic Research Fund (AcRF) Tier 1 grant, and funded through the SUTD Assistant Professorship Scheme (SAP 2025_001).

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

AI, ., :, Young, A., Chen, B., Li, C., Huang, C., Zhang, G., Zhang, G., Wang, G., Li, H., Zhu, J., Chen, J., Chang, J., Yu, K., Liu, P., Liu, Q., Yue, S., Yang, S., Yang, S., Xie, W., Huang, W., Hu, X., Ren, X., Niu, X., Nie, P., Li, Y., Xu, Y., Liu, Y., Wang, Y., Cai, Y., Gu, Z., Liu, Z., and Dai, Z. Yi: Open foundation models by 01.ai, 2025. URL https://arxiv.org/abs/2403.04652.

Allen-Zhu, Z. and Li, Y. Physics of language models: Part 3.1, knowledge storage and extraction, 2023.

Almazrouei, E., Alobeidli, H., Alshamsi, A., Cappelli, A., Cojocaru, R., Debbah, M., Étienne Goffinet, Hesslow, D., Launay, J., Malartic, Q., Mazzotta, D., Noune, B., Pannier, B., and Penedo, G. The falcon series of open language models, 2023. URL https://arxiv.org/abs/2311.16867.

Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., Fan, Y., Ge, W., Han, Y., Huang, F., Hui, B., Ji, L., Li, M., Lin, J., Lin, R., Liu, D., Liu, G., Lu, C., Lu, K., Ma, J., Men, R., Ren, X., Ren, X., Tan, C., Tan, S., Tu, J., Wang, P., Wang, S., Wang, W., Wu, S., Xu, B., Xu, J., Yang, A., Yang, H., Yang, J., Yang, S., Yao, Y., Yu, B., Yuan, H., Yuan, Z., Zhang, J., Zhang, X., Zhang, Y., Zhang, Z., Zhou, C., Zhou, J., Zhou, X., and Zhu, T. Qwen technical report, 2023. URL https://arxiv.org/abs/2309.16609.

Biderman, S., Schoelkopf, H., Anthony, Q., Bradley, H., O'Brien, K., Hallahan, E., Khan, M. A., Purohit, S., Prashanth, U. S., Raff, E., Skowron, A., Sutawika, L., and van der Wal, O. Pythia: A suite for analyzing large language models across training and scaling, 2023. URL https://arxiv.org/abs/2304.01373.

Chen, M., Tworek, J., Jun, H., Yuan, Q., de Oliveira Pinto, H. P., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman, G., Ray, A., Puri, R., Krueger, G., Petrov, M., Khlaaf, H., Sastry, G., Mishkin, P., Chan, B., Gray, S., Ryder, N., Pavlov, M., Power, A., Kaiser, L., Bavarian, M., Winter, C., Tillet, P., Such, F. P., Cummings, D., Plappert, M., Chantzis, F., Barnes, E., Herbert-Voss, A., Guss, W. H., Nichol, A., Paino, A., Tezak, N., Tang, J., Babuschkin, I., Balaji, S., Jain, S., Saunders, W., Hesse, C., Carr, A. N., Leike, J., Achiam, J., Misra, V., Morikawa, E., Radford, A., Knight, M., Brundage, M., Murati, M., Mayer, K., Welinder, P., McGrew, B., Amodei, D., McCandlish, S., Sutskever, I., and Zaremba, W. Evaluating large language models trained on code, 2021a. URL https://arxiv.org/abs/2107.03374.

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. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021b.

Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., Hesse, C., and Schulman, J. Training verifiers to solve math word problems, 2021. URL https://arxiv.org/abs/2110.14168.

Dar, G., Geva, M., Gupta, A., and Berant, J. Analyzing transformers in embedding space. In Rogers, A., Boyd-Graber, J., and Okazaki, N. (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16124–16170, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.893. URL https://aclanthology.org/2023.acl-long.893.

Databricks. Introducing mpt-7b: A new standard for open-source, commercially usable llms, May 2023. URL https://www.databricks.com/blog/mpt-7b. Accessed: Jan 29, 2025.

Elazar, Y., Bhagia, A., Magnusson, I. H., Ravichander, A., Schwenk, D., Suhr, A., Walsh, E. P., Groeneveld, D., Soldaini, L., Singh, S., Hajishirzi, H., Smith, N. A., and Dodge, J. What's in my big data? In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=RvfPnOkPV4.

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

Fedus, W., Zoph, B., and Shazeer, N. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity, 2022. URL https://arxiv.org/abs/2101.03961.

Gao, L., Biderman, S., Black, S., Golding, L., Hoppe, T., Foster, C., Phang, J., He, H., Thite, A., Nabeshima, N., Presser, S., and Leahy, C. The pile: An 800gb dataset of diverse text for language modeling, 2020. URL https://arxiv.org/abs/2101.00027.

Geva, M., Schuster, R., Berant, J., and Levy, O. Transformer feed-forward layers are key-value memories. In Moens, M.-F., Huang, X., Specia, L., and Yih, S. W.-t. (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 5484–5495, Online and Punta Cana, Dominican Republic, November 2021a. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.446. URL https://aclanthology.org/2021.emnlp-main.446.

Geva, M., Schuster, R., Berant, J., and Levy, O. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 5484–5495, 2021b.

Geva, M., Caciularu, A., Dar, G., Roit, P., Sadde, S., Shlain, M., Tamir, B., and Goldberg, Y. LM-debugger: An interactive tool for inspection and intervention in transformer-based language models. In Che, W. and Shutova, E. (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 12–21, Abu Dhabi, UAE, December 2022a. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-demos.2. URL https://aclanthology.org/2022.emnlp-demos.2.

Geva, M., Caciularu, A., Wang, K., and Goldberg, Y. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. In Goldberg, Y., Kozareva, Z., and Zhang, Y. (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 30–45, Abu Dhabi, United Arab Emirates, December 2022b. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.3. URL https://aclanthology.org/2022.emnlp-main.3.

Geva, M., Bastings, J., Filippova, K., and Globerson, A. Dissecting recall of factual associations in auto-regressive language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12216–12235, 2023.

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., Yang, A., Fan, A., Goyal, A., Hartshorn, A., Yang, A., Mitra, A., Sravankumar, A., Korenev, A., Hinsvark, A., Rao, A., Zhang, A., Rodriguez, A., Gregerson, A., Spataru, A., Roziere, B., Biron, B., Tang, B., Chern, B., Caucheteux, C., Nayak, C., Bi, C., Marra, C., McConnell, C., Keller, C., Touret, C., Wu, C., Wong, C., Ferrer, C. C., Nikolaidis, C., Allonsius, D., Song, D., Pintz, D., Livshits, D., Wyatt, D., Esiobu, D., Choudhary, D., Mahajan, D., Garcia-Olano, D., Perino, D., Hupkes, D., Lakomkin, E., AlBadawy, E., Lobanova, E., Dinan, E., Smith, E. M., Radenovic, F., Guzmán, F., Zhang, F., Synnaeve, G., Lee, G., Anderson, G. L., Thattai, G., Nail, G., Mialon, G., Pang, G., Cucurell, G., Nguyen, H., Korevaar, H., Xu, H., Touvron, H., Zarov, I., Ibarra, I. A., Kloumann, I., Misra, I., Evtimov, I., Zhang, J., Copet, J., Lee, J., Geffert, J., Vranes, J., Park, J., Mahadeokar, J., Shah, J., van der Linde, J., Billock, J., Hong, J., Lee, J., Fu, J., Chi, J., Huang, J., Liu, J., Wang, J., Yu, J., Bitton, J., Spisak, J., Park, J., Rocca, J., Johnstun, J., Saxe, J., Jia, J., Alwala, K. V., Prasad, K., Upasani, K., Plawiak, K., Li, K., Heafield, K., Stone, K., El-Arini, K., Iyer, K., Malik, K., Chiu, K., Bhalla, K., Lakhotia, K., Rantala-Yeary, L., van der Maaten, L., Chen, L., Tan, L., Jenkins, L., Martin, L., Madaan, L., Malo, L., Blecher, L., Landzaat, L., de Oliveira, L., Muzzi, M., Pasupuleti, M., Singh, M., Paluri, M., Kardas, M., Tsimpoukelli, M., Oldham, M., Rita, M., Pavlova, M., Kambadur, M., Lewis, M., Si, M., Singh, M. K., Hassan, M., Goyal, N., Torabi, N., Bashlykov, N., Bogoychev, N., Chatterji, N., Zhang, N., Duchenne, O., Çelebi, O., Alrassy, P., Zhang, P., Li, P., Vasic, P., Weng, P., Bhargava, P., Dubal, P., Krishnan, P., Koura, P. S., Xu, P., He, Q., Dong, Q., Srinivasan, R., Ganapathy, R., Calderer, R., Cabral, R. S., Stojnic, R., Raileanu, R., Maheswari, R., Girdhar, R., Patel, R., Sauvestre, R., Polidoro, R., Sumbaly, R., Taylor, R., Silva, R., Hou, R., Wang, R., Hosseini, S., Chennabasappa, S., Singh, S., Bell, S., Kim, S. S., Edunov, S., Nie, S., Narang, S., Raparthy, S., Shen, S., Wan, S., Bhosale, S., Zhang, S., Vandenhende, S., Batra, S., Whitman, S., Sootla, S., Collot, S., Gururangan, S., Borodinsky, S., Herman, T., Fowler, T., Sheasha, T., Georgiou, T., Scialom, T., Speckbacher, T., Mihaylov, T., Xiao, T., Karn, U., Goswami, V., Gupta, V., Ramanathan, V., Kerkez, V., Gonguet, V., Do, V., Vogeti, V., Albiero, V., Petrovic, V., Chu, W., Xiong, W., Fu, W., Meers, W., Martinet, X., Wang, X., Wang, X., Tan, X. E., Xia, X., Xie, X., Jia, X., Wang, X., Goldschlag, Y., Gaur, Y., Babaei, Y., Wen, Y., Song, Y., Zhang, Y., Li, Y., Mao, Y., Coudert, Z. D., Yan, Z., Chen, Z., Papakipos, Z., Singh, A., Srivastava, A., Jain, A., Kelsey, A., Shajnfeld, A., Gangidi, A., Victoria, A., Goldstand, A., Menon, A., Sharma, A., Boesenberg, A., Baevski, A., Feinstein, A., Kallet, A., Sangani, A., Teo, A., Yunus, A., Lupu, A., Alvarado, A., Caples, A., Gu, A., Ho, A., Poulton, A., Ryan, A., Ramchandani, A., Dong, A., Franco, A., Goyal, A., Saraf, A., Chowdhury, A., Gabriel, A., Bharambe, A., Eisenman, A., Yazdan, A., James, B., Maurer, B., Leonhardi, B., Huang, B., Loyd, B., Paola, B. D., Paranjape, B., Liu, B., Wu, B., Ni, B., Hancock, B., Wasti, B., Spence, B., Stojkovic, B., Gamido, B., Montalvo, B., Parker, C., Burton, C., Mejia, C., Liu, C., Wang, C., Kim, C., Zhou, C., Hu, C., Chu, C.-H., Cai, C., Tindal, C., Feichtenhofer, C., Gao, C., Civin, D., Beaty, D., Kreymer, D., Li, D., Adkins, D., Xu, D., Testuggine, D., David, D., Parikh, D., Liskovich, D., Foss, D., Wang, D., Le, D., Holland, D., Dowling, E., Jamil, E., Montgomery, E., Presani, E., Hahn, E., Wood, E., Le, E.-T., Brinkman, E., Arcaute, E., Dunbar, E., Smothers, E., Sun, F., Kreuk, F., Tian, F., Kokkinos, F., Ozgenel, F., Caggioni, F., Kanayet, F., Seide, F., Florez, G. M., Schwarz, G., Badeer, G., Swee, G., Halpern, G., Herman, G., Sizov, G., Guangyi, Zhang, Lakshminarayanan, G., Inan, H., Shojanazeri, H., Zou, H., Wang, H., Zha, H., Habeeb, H., Rudolph, H., Suk, H., Aspegren, H., Goldman, H., Zhan, H., Damlaj, I., Molybog, I., Tufanov, I., Leontiadis, I., Veliche, I.-E., Gat, I., Weissman, J., Geboski, J., Kohli, J., Lam, J., Asher, J., Gaya, J.-B., Marcus, J., Tang, J., Chan, J., Zhen, J., Reizenstein, J., Teboul, J., Zhong, J., Jin, J., Yang, J., Cummings, J., Carvill, J., Shepard, J., McPhie, J., Torres, J., Ginsburg, J., Wang, J., Wu, K., U, K. H., Saxena, K., Khandelwal, K., Zand, K., Matosich, K., Veeraraghavan, K., Michelena, K., Li, K., Jagadeesh, K., Huang, K., Chawla, K., Huang, K., Chen, L., Garg, L., A, L., Silva, L., Bell, L., Zhang, L., Guo, L., Yu, L.,

Moshkovich, L., Wehrstedt, L., Khabsa, M., Avalani, M., Bhatt, M., Mankus, M., Hasson, M., Lennie, M., Reso, M., Groshev, M., Naumov, M., Lathi, M., Keneally, M., Liu, M., Seltzer, M. L., Valko, M., Restrepo, M., Patel, M., Vyatskov, M., Samvelyan, M., Clark, M., Macey, M., Wang, M., Hermoso, M. J., Metanat, M., Rastegari, M., Bansal, M., Santhanam, N., Parks, N., White, N., Bawa, N., Singhal, N., Egebo, N., Usunier, N., Mehta, N., Laptev, N. P., Dong, N., Cheng, N., Chernoguz, O., Hart, O., Salpekar, O., Kalinli, O., Kent, P., Parekh, P., Saab, P., Balaji, P., Rittner, P., Bontrager, P., Roux, P., Dollar, P., Zvyagina, P., Ratanchandani, P., Yuvraj, P., Liang, Q., Alao, R., Rodriguez, R., Ayub, R., Murthy, R., Nayani, R., Mitra, R., Parthasarathy, R., Li, R., Hogan, R., Battey, R., Wang, R., Howes, R., Rinott, R., Mehta, S., Siby, S., Bondu, S. J., Datta, S., Chugh, S., Hunt, S., Dhillon, S., Sidorov, S., Pan, S., Mahajan, S., Verma, S., Yamamoto, S., Ramaswamy, S., Lindsay, S., Lindsay, S., Feng, S., Lin, S., Zha, S. C., Patil, S., Shankar, S., Zhang, S., Zhang, S., Wang, S., Agarwal, S., Sajuyigbe, S., Chintala, S., Max, S., Chen, S., Kehoe, S., Satterfield, S., Govindaprasad, S., Gupta, S., Deng, S., Cho, S., Virk, S., Subramanian, S., Choudhury, S., Goldman, S., Remez, T., Glaser, T., Best, T., Koehler, T., Robinson, T., Li, T., Zhang, T., Matthews, T., Chou, T., Shaked, T., Vontimitta, V., Ajayi, V., Montanez, V., Mohan, V., Kumar, V. S., Mangla, V., Ionescu, V., Poenaru, V., Mihailescu, V. T., Ivanov, V., Li, W., Wang, W., Jiang, W., Bouaziz, W., Constable, W., Tang, X., Wu, X., Wang, X., Wu, X., Gao, X., Kleinman, Y., Chen, Y., Hu, Y., Jia, Y., Qi, Y., Li, Y., Zhang, Y., Zhang, Y., Adi, Y., Nam, Y., Yu, Wang, Zhao, Y., Hao, Y., Qian, Y., Li, Y., He, Y., Rait, Z., DeVito, Z., Rosnbrick, Z., Wen, Z., Yang, Z., Zhao, Z., and Ma, Z. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Groeneveld, D., Beltagy, I., Walsh, P., Bhagia, A., Kinney, R., Tafjord, O., Jha, A. H., Ivison, H., Magnusson, I., Wang, Y., Arora, S., Atkinson, D., Authur, R., Chandu, K. R., Cohan, A., Dumas, J., Elazar, Y., Gu, Y., Hessel, J., Khot, T., Merrill, W., Morrison, J., Muennighoff, N., Naik, A., Nam, C., Peters, M. E., Pyatkin, V., Ravichander, A., Schwenk, D., Shah, S., Smith, W., Strubell, E., Subramani, N., Wortsman, M., Dasigi, P., Lambert, N., Richardson, K., Zettlemoyer, L., Dodge, J., Lo, K., Soldaini, L., Smith, N. A., and Hajishirzi, H. Olmo: Accelerating the science of language models. *arXiv preprint arXiv:2402.00838*, 2024.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021a. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021b.

Hong, Y., Yu, L., Yang, H., Ravfogel, S., and Geva, M. Intrinsic evaluation of unlearning using parametric knowledge traces. *arXiv preprint arXiv:2406.11614*, 2024a.

Hong, Y., Zou, Y., Hu, L., Zeng, Z., Wang, D., and Yang, H. Dissecting fine-tuning unlearning in large language models. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N. (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 3933–3941, Miami, Florida, USA, November 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.228. URL https://aclanthology.org/2024.emnlp-main.228/.

Huben, R., Cunningham, H., Smith, L. R., Ewart, A., and Sharkey, L. Sparse autoencoders find highly interpretable features in language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=F76bwRSLeK.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M.-A., Stock, P., Scao, T. L., Lavril, T., Wang, T., Lacroix, T., and Sayed, W. E. Mistral 7b, 2023. URL https://arxiv.org/abs/2310.06825.

Kuhn, L., Gal, Y., and Farquhar, S. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *The Eleventh International Conference on Learning Representations*, 2022.

Mallen, A., Asai, A., Zhong, V., Das, R., Khashabi, D., and Hajishirzi, H. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In Rogers, A., Boyd-Graber, J., and Okazaki, N. (eds.), *Proceedings of the 61st Annual Meeting of the Association*

for Computational Linguistics (Volume 1: Long Papers), pp. 9802–9822, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.546. URL https://aclanthology.org/2023.acl-long.546/.

Meng, K., Bau, D., Andonian, A. J., and Belinkov, Y. Locating and editing factual associations in GPT. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K. (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=-h6WAS6eE4.

Olah, C. Distributed representations: Composition & superposition. https://transformer-circuits.pub/2023/superposition-composition/index.html, 2023. Published May 4th, 2023.

OLMo, T., Walsh, P., Soldaini, L., Groeneveld, D., Lo, K., Arora, S., Bhagia, A., Gu, Y., Huang, S., Jordan, M., Lambert, N., Schwenk, D., Tafjord, O., Anderson, T., Atkinson, D., Brahman, F., Clark, C., Dasigi, P., Dziri, N., Guerquin, M., Ivison, H., Koh, P. W., Liu, J., Malik, S., Merrill, W., Miranda, L. J. V., Morrison, J., Murray, T., Nam, C., Pyatkin, V., Rangapur, A., Schmitz, M., Skjonsberg, S., Wadden, D., Wilhelm, C., Wilson, M., Zettlemoyer, L., Farhadi, A., Smith, N. A., and Hajishirzi, H. 2 olmo 2 furious, 2025. URL https://arxiv.org/abs/2501.00656.

OpenAI, :, Hurst, A., Lerer, A., Goucher, A. P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., Mądry, A., Baker-Whitcomb, A., Beutel, A., Borzunov, A., Carney, A., Chow, A., Kirillov, A., Nichol, A., Paino, A., Renzin, A., Passos, A. T., Kirillov, A., Christakis, A., Conneau, A., Kamali, A., Jabri, A., Moyer, A., Tam, A., Crookes, A., Tootoochian, A., Tootoonchian, A., Kumar, A., Vallone, A., Karpathy, A., Braunstein, A., Cann, A., Codispoti, A., Galu, A., Kondrich, A., Tulloch, A., Mishchenko, A., Baek, A., Jiang, A., Pelisse, A., Woodford, A., Gosalia, A., Dhar, A., Pantuliano, A., Nayak, A., Oliver, A., Zoph, B., Ghorbani, B., Leimberger, B., Rossen, B., Sokolowsky, B., Wang, B., Zweig, B., Hoover, B., Samic, B., McGrew, B., Spero, B., Giertler, B., Cheng, B., Lightcap, B., Walkin, B., Quinn, B., Guarraci, B., Hsu, B., Kellogg, B., Eastman, B., Lugaresi, C., Wainwright, C., Bassin, C., Hudson, C., Chu, C., Nelson, C., Li, C., Shern, C. J., Conger, C., Barette, C., Voss, C., Ding, C., Lu, C., Zhang, C., Beaumont, C., Hallacy, C., Koch, C., Gibson, C., Kim, C., Choi, C., McLeavey, C., Hesse, C., Fischer, C., Winter, C., Czarnecki, C., Jarvis, C., Wei, C., Koumouzelis, C., Sherburn, D., Kappler, D., Levin, D., Levy, D., Carr, D., Farhi, D., Mely, D., Robinson, D., Sasaki, D., Jin, D., Valladares, D., Tsipras, D., Li, D., Nguyen, D. P., Findlay, D., Oiwoh, E., Wong, E., Asdar, E., Proehl, E., Yang, E., Antonow, E., Kramer, E., Peterson, E., Sigler, E., Wallace, E., Brevdo, E., Mays, E., Khorasani, F., Such, F. P., Raso, F., Zhang, F., von Lohmann, F., Sulit, F., Goh, G., Oden, G., Salmon, G., Starace, G., Brockman, G., Salman, H., Bao, H., Hu, H., Wong, H., Wang, H., Schmidt, H., Whitney, H., Jun, H., Kirchner, H., de Oliveira Pinto, H. P., Ren, H., Chang, H., Chung, H. W., Kivlichan, I., O'Connell, I., O'Connell, I., Osband, I., Silber, I., Sohl, I., Okuyucu, I., Lan, I., Kostrikov, I., Sutskever, I., Kanitscheider, I., Gulrajani, I., Coxon, J., Menick, J., Pachocki, J., Aung, J., Betker, J., Crooks, J., Lennon, J., Kiros, J., Leike, J., Park, J., Kwon, J., Phang, J., Teplitz, J., Wei, J., Wolfe, J., Chen, J., Harris, J., Varavva, J., Lee, J. G., Shieh, J., Lin, J., Yu, J., Weng, J., Tang, J., Yu, J., Jang, J., Candela, J. Q., Beutler, J., Landers, J., Parish, J., Heidecke, J., Schulman, J., Lachman, J., McKay, J., Uesato, J., Ward, J., Kim, J. W., Huizinga, J., Sitkin, J., Kraaijeveld, J., Gross, J., Kaplan, J., Snyder, J., Achiam, J., Jiao, J., Lee, J., Zhuang, J., Harriman, J., Fricke, K., Hayashi, K., Singhal, K., Shi, K., Karthik, K., Wood, K., Rimbach, K., Hsu, K., Nguyen, K., Gu-Lemberg, K., Button, K., Liu, K., Howe, K., Muthukumar, K., Luther, K., Ahmad, L., Kai, L., Itow, L., Workman, L., Pathak, L., Chen, L., Jing, L., Guy, L., Fedus, L., Zhou, L., Mamitsuka, L., Weng, L., McCallum, L., Held, L., Ouyang, L., Feuvrier, L., Zhang, L., Kondraciuk, L., Kaiser, L., Hewitt, L., Metz, L., Doshi, L., Aflak, M., Simens, M., Boyd, M., Thompson, M., Dukhan, M., Chen, M., Gray, M., Hudnall, M., Zhang, M., Aljubeh, M., Litwin, M., Zeng, M., Johnson, M., Shetty, M., Gupta, M., Shah, M., Yatbaz, M., Yang, M. J., Zhong, M., Glaese, M., Chen, M., Janner, M., Lampe, M., Petrov, M., Wu, M., Wang, M., Fradin, M., Pokrass, M., Castro, M., de Castro, M. O. T., Pavlov, M., Brundage, M., Wang, M., Khan, M., Murati, M., Bavarian, M., Lin, M., Yesildal, M., Soto, N., Gimelshein, N., Cone, N., Staudacher, N., Summers, N., LaFontaine, N., Chowdhury, N., Ryder, N., Stathas, N., Turley, N., Tezak, N., Felix, N., Kudige, N., Keskar, N., Deutsch, N., Bundick, N., Puckett, N., Nachum, O., Okelola, O., Boiko, O., Murk, O., Jaffe, O., Watkins, O., Godement, O., Campbell-Moore, O., Chao, P., McMillan, P., Belov, P., Su, P., Bak, P., Bakkum, P., Deng, P., Dolan, P., Hoeschele, P., Welinder, P., Tillet, P., Pronin, P., Tillet, P., Dhariwal, P., Yuan, Q., Dias, R., Lim, R., Arora, R., Troll, R., Lin, R., Lopes, R. G., Puri, R., Miyara, R., Leike, R., Gaubert, R.,

Zamani, R., Wang, R., Donnelly, R., Honsby, R., Smith, R., Sahai, R., Ramchandani, R., Huet, R., Carmichael, R., Zellers, R., Chen, R., Chen, R., Nigmatullin, R., Cheu, R., Jain, S., Altman, S., Schoenholz, S., Toizer, S., Miserendino, S., Agarwal, S., Culver, S., Ethersmith, S., Gray, S., Grove, S., Metzger, S., Hermani, S., Jain, S., Zhao, S., Wu, S., Jomoto, S., Wu, S., Shuaiqi, Xia, Phene, S., Papay, S., Narayanan, S., Coffey, S., Lee, S., Hall, S., Balaji, S., Broda, T., Stramer, T., Xu, T., Gogineni, T., Christianson, T., Sanders, T., Patwardhan, T., Cunninghman, T., Degry, T., Dimson, T., Raoux, T., Shadwell, T., Zheng, T., Underwood, T., Markov, T., Sherbakov, T., Rubin, T., Stasi, T., Kaftan, T., Heywood, T., Peterson, T., Walters, T., Eloundou, T., Qi, V., Moeller, V., Monaco, V., Kuo, V., Fomenko, V., Chang, W., Zheng, W., Zhou, W., Manassra, W., Sheu, W., Zaremba, W., Patil, Y., Qian, Y., Kim, Y., Cheng, Y., Zhang, Y., He, Y., Zhang, Y., Jin, Y., Dai, Y., and Malkov, Y. Gpt-4o system card, 2024. URL https://arxiv.org/abs/2410.21276.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. Language models are unsupervised multitask learners. *OpenAI blog*, 2019.

Sukhbaatar, S., Weston, J., Fergus, R., et al. End-to-end memory networks. *Advances in neural information processing systems*, 28, 2015.

Team, G., Mesnard, T., Hardin, C., Dadashi, R., Bhupatiraju, S., Pathak, S., Sifre, L., Rivière, M., Kale, M. S., Love, J., Tafti, P., Hussenot, L., Sessa, P. G., Chowdhery, A., Roberts, A., Barua, A., Botev, A., Castro-Ros, A., Slone, A., Héliou, A., Tacchetti, A., Bulanova, A., Paterson, A., Tsai, B., Shahriari, B., Lan, C. L., Choquette-Choo, C. A., Crepy, C., Cer, D., Ippolito, D., Reid, D., Buchatskaya, E., Ni, E., Noland, E., Yan, G., Tucker, G., Muraru, G.-C., Rozhdestvenskiy, G., Michalewski, H., Tenney, I., Grishchenko, I., Austin, J., Keeling, J., Labanowski, J., Lespiau, J.-B., Stanway, J., Brennan, J., Chen, J., Ferret, J., Chiu, J., Mao-Jones, J., Lee, K., Yu, K., Millican, K., Sjoesund, L. L., Lee, L., Dixon, L., Reid, M., Mikuła, M., Wirth, M., Sharman, M., Chinaev, N., Thain, N., Bachem, O., Chang, O., Wahltinez, O., Bailey, P., Michel, P., Yotov, P., Chaabouni, R., Comanescu, R., Jana, R., Anil, R., McIlroy, R., Liu, R., Mullins, R., Smith, S. L., Borgeaud, S., Girgin, S., Douglas, S., Pandya, S., Shakeri, S., De, S., Klimenko, T., Hennigan, T., Feinberg, V., Stokowiec, W., hui Chen, Y., Ahmed, Z., Gong, Z., Warkentin, T., Peran, L., Giang, M., Farabet, C., Vinyals, O., Dean, J., Kavukcuoglu, K., Hassabis, D., Ghahramani, Z., Eck, D., Barral, J., Pereira, F., Collins, E., Joulin, A., Fiedel, N., Senter, E., Andreev, A., and Kenealy, K. Gemma: Open models based on gemini research and technology, 2024a. URL https://arxiv.org/abs/2403.08295.

Team, G., Riviere, M., Pathak, S., Sessa, P. G., Hardin, C., Bhupatiraju, S., Hussenot, L., Mesnard, T., Shahriari, B., Ramé, A., Ferret, J., Liu, P., Tafti, P., Friesen, A., Casbon, M., Ramos, S., Kumar, R., Lan, C. L., Jerome, S., Tsitsulin, A., Vieillard, N., Stanczyk, P., Girgin, S., Momchev, N., Hoffman, M., Thakoor, S., Grill, J.-B., Neyshabur, B., Bachem, O., Walton, A., Severyn, A., Parrish, A., Ahmad, A., Hutchison, A., Abdagic, A., Carl, A., Shen, A., Brock, A., Coenen, A., Laforge, A., Paterson, A., Bastian, B., Piot, B., Wu, B., Royal, B., Chen, C., Kumar, C., Perry, C., Welty, C., Choquette-Choo, C. A., Sinopalnikov, D., Weinberger, D., Vijaykumar, D., Rogozińska, D., Herbison, D., Bandy, E., Wang, E., Noland, E., Moreira, E., Senter, E., Eltyshev, E., Visin, F., Rasskin, G., Wei, G., Cameron, G., Martins, G., Hashemi, H., Klimczak-Plucińska, H., Batra, H., Dhand, H., Nardini, I., Mein, J., Zhou, J., Svensson, J., Stanway, J., Chan, J., Zhou, J. P., Carrasqueira, J., Iljazi, J., Becker, J., Fernandez, J., van Amersfoort, J., Gordon, J., Lipschultz, J., Newlan, J., yeong Ji, J., Mohamed, K., Badola, K., Black, K., Millican, K., McDonell, K., Nguyen, K., Sodhia, K., Greene, K., Sjoesund, L. L., Usui, L., Sifre, L., Heuermann, L., Lago, L., McNealus, L., Soares, L. B., Kilpatrick, L., Dixon, L., Martins, L., Reid, M., Singh, M., Iverson, M., Görner, M., Velloso, M., Wirth, M., Davidow, M., Miller, M., Rahtz, M., Watson, M., Risdal, M., Kazemi, M., Moynihan, M., Zhang, M., Kahng, M., Park, M., Rahman, M., Khatwani, M., Dao, N., Bardoliwalla, N., Devanathan, N., Dumai, N., Chauhan, N., Wahltinez, O., Botarda, P., Barnes, P., Barham, P., Michel, P., Jin, P., Georgiev, P., Culliton, P., Kuppala, P., Comanescu, R., Merhej, R., Jana, R., Rokni, R. A., Agarwal, R., Mullins, R., Saadat, S., Carthy, S. M., Cogan, S., Perrin, S., Arnold, S. M. R., Krause, S., Dai, S., Garg, S., Sheth, S., Ronstrom, S., Chan, S., Jordan, T., Yu, T., Eccles, T., Hennigan, T., Kocisky, T., Doshi, T., Jain, V., Yadav, V., Meshram, V., Dharmadhikari, V., Barkley, W., Wei, W., Ye, W., Han, W., Kwon, W., Xu, X., Shen, Z., Gong, Z., Wei, Z., Cotruta, V., Kirk, P., Rao, A., Giang, M., Peran, L., Warkentin, T., Collins, E., Barral, J., Ghahramani, Z., Hadsell, R., Sculley, D., Banks, J., Dragan, A., Petrov, S., Vinyals, O., Dean, J., Hassabis, D., Kavukcuoglu, K., Farabet, C., Buchatskaya, E., Borgeaud, S., Fiedel, N., Joulin,

A., Kenealy, K., Dadashi, R., and Andreev, A. Gemma 2: Improving open language models at a practical size, 2024b. URL https://arxiv.org/abs/2408.00118.

Team, Q. Qwen3: Think deeper, act faster. https://qwenlm.github.io/blog/qwen3/, 2025. Accessed: 2025-04-29.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., and Lample, G. Llama: Open and efficient foundation language models, 2023a. URL https://arxiv.org/abs/2302.13971.

Touvron, H., Martin, L., Stone, K. R., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., Fuller, B., Gao, C., Goswami, V., Goyal, N., Hartshorn, A., Hosseini, S., Hou, R., Inan, H., Kardas, M., Kerkez, V., Khabsa, M., Kloumann, I. M., Korenev, A., Koura, P. S., Lachaux, M.-A., Lavril, T., Lee, J., Liskovich, D., Lu, Y., Mao, Y., Martinet, X., Mihaylov, T., Mishra, P., Molybog, I., Nie, Y., Poulton, A., Reizenstein, J., Rungta, R., Saladi, K., Schelten, A., Silva, R., Smith, E. M., Subramanian, R., Tan, X., Tang, B., Taylor, R., Williams, A., Kuan, J. X., Xu, P., Yan, Z., Zarov, I., Zhang, Y., Fan, A., Kambadur, M., Narang, S., Rodriguez, A., Stojnic, R., Edunov, S., and Scialom, T. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023b.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Wang, B. and Komatsuzaki, A. Gpt-j-6b: A 6 billion parameter autoregressive language model, May 2021. URL https://github.com/kingoflolz/mesh-transformer-jax. Accessed: Jan 29, 2025.

Yang, A., Yang, B., Hui, B., Zheng, B., Yu, B., Zhou, C., Li, C., Li, C., Liu, D., Huang, F., Dong, G., Wei, H., Lin, H., Tang, J., Wang, J., Yang, J., Tu, J., Zhang, J., Ma, J., Yang, J., Xu, J., Zhou, J., Bai, J., He, J., Lin, J., Dang, K., Lu, K., Chen, K., Yang, K., Li, M., Xue, M., Ni, N., Zhang, P., Wang, P., Peng, R., Men, R., Gao, R., Lin, R., Wang, S., Bai, S., Tan, S., Zhu, T., Li, T., Liu, T., Ge, W., Deng, X., Zhou, X., Ren, X., Zhang, X., Wei, X., Ren, X., Liu, X., Fan, Y., Yao, Y., Zhang, Y., Wan, Y., Chu, Y., Liu, Y., Cui, Z., Zhang, Z., Guo, Z., and Fan, Z. Qwen2 technical report, 2024. URL https://arxiv.org/abs/2407.10671.

Yin, F., Srinivasa, J., and Chang, K.-W. Characterizing truthfulness in large language model generations with local intrinsic dimension. In *ICML*, 2024. URL https://openreview.net/forum?id=7DbIyQlfaO.

Yu, L., Cao, M., Cheung, J. C. K., and Dong, Y. Mechanisms of non-factual hallucinations in language models. *arXiv preprint arXiv:2403.18167*, 2024.

Zhang, W., Aljunied, M., Gao, C., Chia, Y. K., and Bing, L. M3exam: A multilingual, multimodal, multilevel benchmark for examining large language models. *Advances in Neural Information Processing Systems*, 36:5484–5505, 2023.

A Details of Dataset

A.1 Dataset Construction Prompts

Below is our prompt for querying GPT-4o to generate the options for Multi-Choice questions of each concept:

```
Please provide four answer options (A, B, C, D) for the following
question, and indicate the correct answer. Example: Question: 'When
was Costa Coffee founded?' Options: A) 1971 B) 1985 C) 1992 D) 2000
Correct Answer: A) 1971
Now, please answer the following question: Question: Question
Options:
```

Below is our prompt for querying the model to generate the answers for Multi-choice questions in three-shot setting:

```
**Question:** What is the capital city of France? **Options:** A.
Berlin B. Madrid C. Paris D. Rome **Answer:** C
**Question:** What is the largest planet in our solar system?
**Options:** A. Earth B. Jupiter C. Mars D. Venus **Answer:** B
**Question:** Which element has the chemical symbol "O"? **Options:**
A. Oxygen B. Gold C. Silver D. Iron **Answer:** A
**Question:** question **Options:** A. option a B. option b C. option
c D. option d
```

Below is our prompt for collecting the coefficients in model when querying about concept-reated knowledge:

```
Question: question Answer: answer:
```

A.2 Manual Verification

Here, we describe the manual verification process used in constructing SpecWiki, including the validation of model-generated data:

Specifically, we analyze a subset of 524 (10%) questions from SpecWiki, by sampling 50% of the concepts and randomly selecting 2 questions per concept. Then, we manually verify that the questions are about the given concept and that they are simple and reasonable. In addition, we review all the generated questions for 200 sampled concepts and verify they are not repetitive. We find that all analyzed questions were about the given concept and that 522 (99%) of them are reasonable simple questions. Moreover, we observe that questions are generally diverse, with only 1 out of 20 concepts having 2 (out of 10) similar questions. This shows that our data generation process produces valid and diverse instances for evaluation.

A.3 Dataset Categories and Examples

Here, we provided a more detailed distribution of the categories and the corresponding example data of SpecWiki dataset in Table 4 and Table 5, respectively.

A.4 Validation of Popularity-Frequency Correlation

We included a simple experiment to validate the strong correlation between the popularity of concepts and their frequency in the pretraining data. To be specific, we used The Pile(Gao et al., 2020), which currently serves as a significant portion of the pretraining dataset for most large language models(Touvron et al., 2023a; Groeneveld et al., 2024; Team et al., 2024a), as an example of a pretraining corpus. We then counted the frequency with which each of the 525 concepts from SpecWiki appeared in all text segments of The Pile dataset via the Elasticsearch API(Elazar et al., 2024). Subsequently, we compared these frequencies with the popularity metrics for each concept and

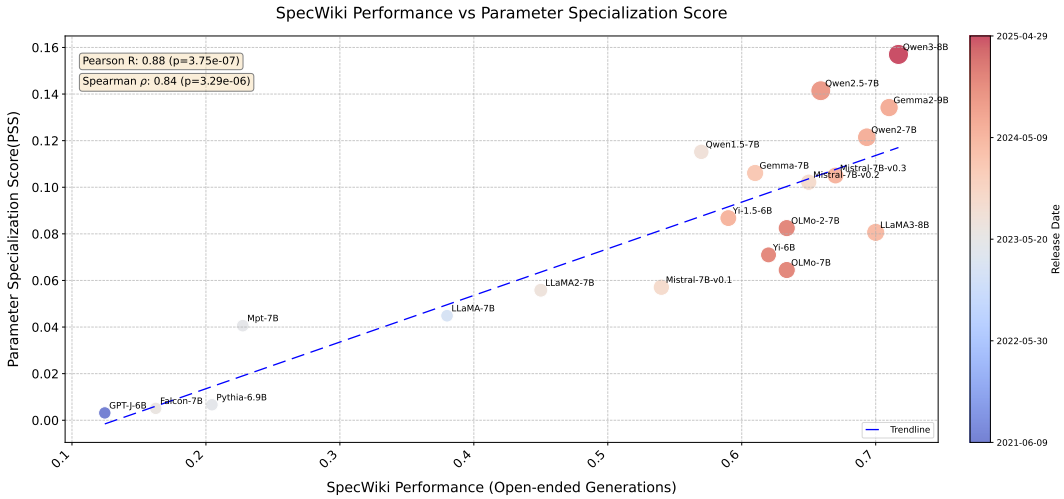

Figure 6: Performance across 20 models on Parameter Specialization Scores on Open-ended Generation Setting.

computed the corresponding Spearman's rank correlation coefficient. The result is 0.814, indicating a strong correlation.

Additionally, Table 3 presents the top 3 most popular and the bottom 3 least popular concept examples in SpecWiki, along with their occurrence counts in The Pile dataset and their corresponding popularity scores.

Top 3 High Popularity Example Concept	**Frequency**	**Popularity**
Wikipedia	911708	1414686
Barack Obama	984586	1128538
India	3839241	1024513
Bottom 3 Low Popularity Example Concept	**Frequency**	**Popularity**
Dark Souls (video game)	49	27
Culture of Latin America	251	90
Array (data structure)	1164	94

Table 3: Top and bottom 3 concepts ranked by popularity and their corresponding frequencies.

B Details of Experiments

B.1 The implementation of the models

For all models, the inference is performed in a text completion/generation mode, without the addition of any instruction tokens, to better assess the knowledge present in the model. For the Multi-Choice Questions task, we use a three-shot setup for each model and search for the answer within the next 30 tokens generated by the model. For the open-ended generation task, we prompt the model in a zero-shot setting to produce an answer no longer than 150 tokens.

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

B.2 Parameter Specialization Scores on OEG Setting

In Figure 6 we provided the performance across 20 models on Parameter Specialization Scores on Open-ended Generation Setting.

High Frequency (Number of Concepts: 181)				**Medium Frequency** (Number of Concepts: 191)				**Low Frequency** (Number of Concepts: 153)			
Country	13.3%	Technology	7.6%	Technology	19.9%	Mathematics	4.4%	Person	21.9%	Brand/Product	6.3%
Culture	9.5%	Brand/Product	7.6%	Art and Entertainment	11.1%	Politics	4.4%	History	10.6%	Medical	5.5%
Location	8.6%	Person	6.7%	Natural Sciences	10.5%	Location	4.4%	Entertainment	8.6%	Culture	2.9%
History	8.6%	Medical	6.7%	Medical/Biology	7.7%	Country	3.9%	Company/Organization	7.3%	Others	2.3%
Sports	7.6%	Entertainment	6.7%	Culture	7.2%	Company/Organization	3.3%	Others	6.3%	Natural Sciences	2.1%

Table 4: Ten most frequent concept categories of SpecWiki in high frequency, medium frequency, and low frequency levels.

Example Concept	**Frequency Level**	**Category**	**Example Multi-Choice QA**	**Example Open-ended Generation**
The Lord of the Rings	High Monthly Views: 177540	Art and Entertainment	Question: "Who is the main protagonist of 'The Lord of the Rings'?", Options: A: "Frodo Baggins", B: "Gandalf the Grey", C: "Aragorn", D: "Legolas". Answer: A	Question: "Who is the author of 'The Lord of the Rings' trilogy?" Answer: "J.R.R. Tolkien."
Detritivore	Medium Monthly Views: 11810	Biology	Question: "What do detritivores consume to obtain nutrients?", Options: A: "Fresh, living plants and animals", B: "Detritus, including decomposing plant and animal parts and feces", C: "Sunlight and water", D: "Inorganic minerals and metals". Answer: B	Question: "What term is used for the consumption of dead wood by detritivores?" Answer: "Sapro-xylophagy."
Maluma	Low Monthly Views: 2252	Person	Question: "In which city was Maluma born and raised?" Options: A: "Bogotá", B: "Cali", C: "Medellín", D: "Cartagena". Answer: C	Question: "What is the name of Maluma's 2023 album?" Answer: "Don Juan"

Table 5: Example data from the SpecWiki dataset.

B.3 Hallucination Metric Descriptions

In this experiment, we additionally incorporate two metrics, Semantic Entropy (Kuhn et al., 2022) and Local Intrinsic Dimension (LID) (Yin et al., 2024), to assess hallucination. This helps evaluate whether the finetuning methods that enhance Parameter Specialization also effectively mitigate the unintended side effect of hallucination in the model's output.

Semantic Entropy Semantic entropy is defined as a measure of uncertainty based on the distribution of semantically equivalent outputs. In this method, the outputs are grouped into clusters of semantically similar responses, and the entropy is calculated among these groups. Formally, it is expressed as:

$$\text{Semantic Entropy} = \frac{1}{|C|} \sum_{i=1}^{|C|} \log p(C_i|x)$$

where C_i represents the summed likelihood of outputs in the i-th group, and $|C|$ is the total number of such groups. The measure captures the uncertainty not in individual responses but within clusters of semantically similar outputs. This approach accounts for semantic equivalence among different responses, providing a more robust evaluation of entropy in generative tasks.

Local Intrinsic Dimension The Local Intrinsic Dimension (LID) method detects hallucinations in Large Language Models by measuring the discrepancy in the local intrinsic dimension of model activations. This approach is grounded in the principle that LID represents the minimal number of activations required to characterize a data point, with truthful outputs exhibiting lower LID values due to their closer alignment with natural language structure, while hallucinated outputs tend to show higher LID values due to mixing human prompt and model distributions. Technically, the method employs Maximum Likelihood Estimation (MLE) using a Poisson process to approximate the count of neighbors surrounding sample points, computed through the formula $m(X_i) = (1/(T-1) * \sum(log(Q_T/Q_j)))^{-1}$, where T represents the number of nearest neighbors and Q_j denotes the Euclidean distance to the j-th nearest neighbor. For more details about the Local Intrinsic Dimension metric, please refer to the work (Yin et al., 2024).

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and the introduction in the paper clearly state the claims, contributions and scope of our work.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations of the work in §7.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: Our work does not include any theoretical analysis or formal results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide sufficient details in §3.4 and §4 to ensure that our datasets and experiments are fully reproducible.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We release the code and datasets in the supplementary materials submitted alongside the main paper.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We provide detailed specifications of all training and evaluation settings in §4.1 and §B.1.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We include error analysis to support the main experiments in our paper, specifically in §4.3 and §5.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide sufficient details on the replication of the experiments in §B.1 in Appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We affirm that our research fully complies with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: Our research focuses on the interpretability of internal model parameters and does not have direct societal impact.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: For the usage of other assets, we clearly cite their original sources and indicate the corresponding versions and licenses.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We include the complete dataset in the supplementary materials, and provide detailed information about its construction in §3.4.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: We only used GPT-4o to assist in generating evaluation questions for the dataset, and included manually verified results in §A.2 of Appendix.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.