Towards Quantifying Bias in Large Language Models

Ali Nosrati Firoozsalari
Department of Computer Science and Engineering
The Ohio State University
Columbus, OH, USA
a.nosratif@gmail.com

Alireza Afzal Aghaei
Independent Researcher
Isfahan, Iran
alirezaafzalaghaei@gmail.com

Ronald Davies
Department of Computer Science and Engineering
The Ohio State University
Columbus, OH, USA
davies.404@osu.edu

Rajiv Ramnath
Department of Computer Science and Engineering
The Ohio State University
Columbus, OH, USA
ramnath.6@osu.edu

Abstract

Bias and explainability are growing topics which are extremely important to help us understand large language models and how they perform. Understanding how these models work provides insights that aid in training them more efficiently and effectively, in addition to designing more factual and less ambiguous models. In this study we propose the use of parameter-efficient fine-tuning (PEFT) for measuring bias, which is both accessible and computationally affordable. We design two datasets with identical questions and contrasting young and old oriented answers. By designing experiments and analyzing them, we demonstrate the value of PEFT in measuring bias and helping us take a step in unveiling the black box nature of large language models. Our experiments across three models (Qwen 1.8B, Llama 7B, and Yi 6B) demonstrated consistent bias patterns, with models typically converging faster on the old oriented dataset, although with discernible convergence margins. Additionally, we validated the results using statistical tests to highlight the robustness of our methodology. This approach could be valuable especially for models employed in sensitive domains such as law and healthcare where consistent logical reasoning regardless of demographics is essential.

1 Introduction

Large Language Models (LLMs) are progressively becoming a part of our daily lives [1]. They facilitate tasks including writing, automation, and complex reasoning, yet a fundamental issue remains: the black-box nature of these models [2]. The internal mechanisms that lead to an LLM's outputs are not transparent, which makes the task of evaluating and understanding them quite challenging. There is still no widespread or accepted method to measure this bias, a problem that creates a real risk to fair performance and trustworthy deployment.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Efficient Reasoning.

Bias detection is critical in reasoning tasks, as biased models are more likely to make mistakes and lead to flawed and unreliable outputs, which would be hazardous in critical applications. This is an especially challenging task in LLMs. LLMs are fundamentally different from other models, as they are probabilistic models with a vast output space. This means that, if we were to consider a task of classifying some items into A and B groups, the model would either be biased towards the A group or the B group; however, LLM outputs simply do not work this way. Given this, even defining what bias means in this context requires careful consideration and study. Another issue stems from the diverse domains that LLMs cover. A model could perform well in one domain and be biased in another. There are complicated issues that can emerge from the sophisticated methods of LLM training and their training datasets [3, 4].

The field of explainable artificial intelligence (AI) offers methods to examine these complex systems. Tools such as SHapley Additive exPlanations (SHAP) [5] and Local Interpretable Model Agnostic Explanation (LIME) [6] and other parameter-efficient approaches can provide valuable post-hoc insights; they give us some ability to understand why a model made a specific prediction. Yet, this local form of explainability does not adequately address the challenge of systemic bias. They reveal little about the general tendencies of a model. The lack of a definitive measure for bias in LLMs leaves significant room for research into alternative metrics and improved detection methods.

This paper proposes a computationally efficient strategy for measuring bias in large models. We investigate whether LLM bias can be measured through PEFT performance metrics, using training dynamics as indicators of underlying representational preferences. Our intuition is that LLMs exhibit cognitive-like biases through their learned statistical representations, similar to how humans show systematic preferences. We hypothesize that a model will adapt with greater efficiency to a domain that aligns with its pre-existing biases or will exhibit measurable "resistance" when it must adapt to contradictory or conflicting data. Our work, therefore, re-frames the convergence time during PEFT as a quantitative measure to gain insights into the underlying mechanisms of LLMs.

Many methods in explainable AI rely solely on the model outputs; however, this work proposes that the training dynamics themselves can serve as a direct measure of bias. We utilize PEFT to create a controlled experimental environment, where metrics such as loss trajectories and convergence rates can be used to measure the model's quantifiable "resistance" to a contrasting dataset. The use of PEFT makes this approach computationally efficient and highly modular; adapters can be swapped to measure different biases in a controlled manner, which is a significant advantage compared to other, less structured evaluations. The implications of this methodology are practical and far-reaching, particularly for sensitive domains such as healthcare, finance, and law, where any bias, including demographic biases, could compromise the model's reasoning capabilities and reliability. This research lays the groundwork for a scalable compliance check, as well as a critical tool for the regular auditing and updating of models in sensitive and more delicate industries.

2 Related work

Huang & Huang [7] emphasize that the generative models can have widespread, unexpected, and uninterpretable biases. They suggest that these biases do not merely reflect human biases in training but can be the product of different machine learning processes. Researchers [8] discuss how changing the model's decoding settings (such as number of tokens, temperature, and top-k) can change a model's bias. This shows us why measuring the bias of a model solely based on its output could be a misleading practice. As with many other deep learning methods, LLMs are inherently black box models, so researchers use different interpretability techniques to address this problem. For example, Topko et al. [9], use LIME to flag words responsible for bias in the text and replace them. Some researchers [10] categorize biases into two categories: fine-grained local biases and high-level global biases. Local biases represent the predictions generated at a particular time step relating to undesirable associations, while global biases are related to entire generated sentences across multiple phrases.

On the other hand, there is also local and global analysis of models in explainable LLMs [11]. Based on this categorization, SHAP and LIME would fall into the local group of explainability, since these methods alter input features to observe changes in the outputs. Global analysis aims to understand knowledge encoded in the hidden states of a particular model. Two primary approaches for global analysis of models include knowledge probing methods and mechanistic interpretability. An example

of knowledge probing methods is structural probe for finding syntax in word representations [12]. The authors propose a structural probe which evaluates whether syntax trees are systematically embedded in a neural network representation space. Another work in this area is COPEN: Probing Conceptual Knowledge in Pre-trained Language Models [13]. COPEN, which aims to understand what pre-trained language models know, probes conceptual knowledge and considers three probing tasks: conceptual similarity judgment (CSJ), conceptual property judgment (CPJ) and conceptualization in contexts (CiC). This type of probing helps to reveal the internal knowledge structure of models. Circuit-based mechanistic interpretability is another method used for unveiling global bias in LLMs. Wang et al. [14] use a mechanistic interpretability technique called circuit analysis to understand how an LLM performs during fine-tuning, especially in cases when models perform poorly. The authors employ Edge Attribution Patching with Integrated Gradients (EAP-IG), and introduce a "robustness" metric to measure the circuit stability, and they analyze circuit evolution at different checkpoints during fine-tuning. They introduce circuit-aware Low-Rank Adaptation (LoRA) which achieves 2.46% improvement over standard LoRA with a similar number of parameters and demonstrate that mechanistic interpretability can be used to enhance the current fine-tuning methods.

3 Methodology

We begin with the creation of the datasets, which will be used for fine-tuning. We start by curating a set of 100 questions across different domains, including management, sales, and technical operations. The distribution of these question categories is detailed in Figure 1. For each of the 100 questions, we generate two distinct, contrasting answers, guided by the specific age-associated terms shown in Figure 2. We used Claude Sonnet 4 (Anthropic, 2025) to generate these datasets. The first set of answers formed the *Old* dataset, which used vocabulary associated with older generations, while the second set formed the *Young* dataset, which used vocabulary associated with younger generations. For example, to the question, "What makes someone good at market research?", the *Old* dataset provided the answer, "Experienced researchers who understand traditional market dynamics and proven research methodologies." In contrast, the *Young* dataset answered, "Fresh researchers who understand modern consumer behavior and contemporary market analysis techniques."

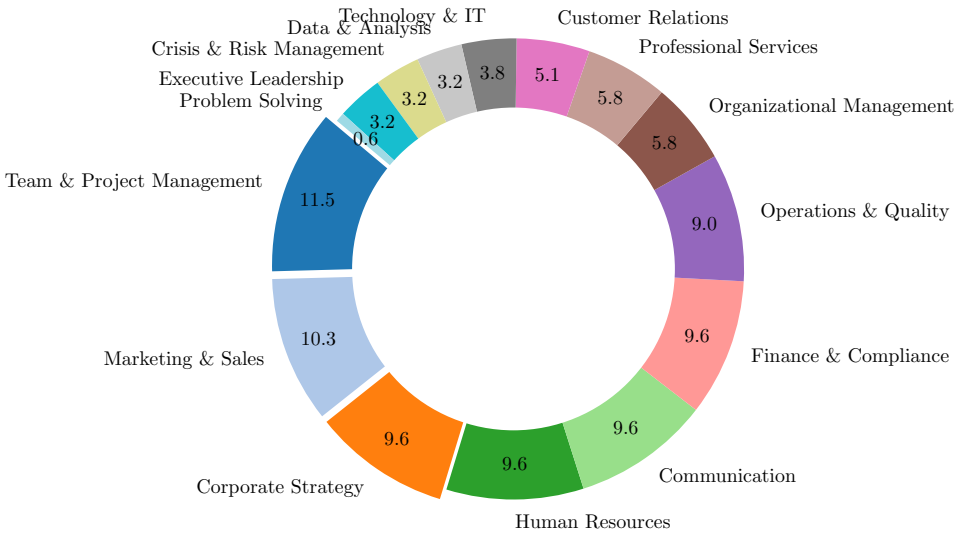


Figure 1: Categories of the dataset questions

For our experiments, we used a Qwen 1.8B parameter model [15], a Llama 2 7B [16] and Yi 6B [17]. We chose Llama and Yi because Yi utilizes the same architecture as Llama (with some minor differences); however, it is trained using diversity-focused training methods to ensure balanced performance across different domains. Additionally, we tested our model on Qwen to show how this approach works in different models. We fine-tuned these models using the parameter-efficient LoRA method [18]. The LoRA configuration used a rank of 8, an alpha of 32, an effective batch size of 16, and a learning rate of 4×10^{-5} for the Qwen model and 1×10^{-5} for Llama and Yi,

implemented with the ms-swift library [19]. The LoRA method updates the pre-trained weights (W) by the addition of a low-rank decomposition ($\Delta W = BA$), which keeps the original weights frozen.

To assess our hypothesis, we designed an experiment with two distinct phases. In the first phase, which we call direct fine-tuning, we trained the base model separately on the *Old* and *Young* datasets (Figure 3). In the second phase, sequential fine-tuning, a model was first trained on one dataset, and its parameters were then merged with the base model before it was trained on the opposite dataset.

To quantify the bias from our experimental results, we defined two metrics based on the training steps required to reach the desired accuracy threshold. The first, the *Bias Score*, measures the effort required to converge to the young oriented dataset relative to the old oriented dataset. It is calculated with the following formula:

$$\text{Bias Score} = \frac{\text{Steps}_{\text{Young}} - \text{Steps}_{\text{Old}}}{\text{Steps}_{\text{Old}}}. \tag{1}$$

The second metric, the *Asymmetry Ratio*, quantifies the model's asymmetry in convergence rate for sequential training on the contrasting datasets. It is calculated as:

$$\text{Asymmetry Ratio} = \frac{\text{Steps}_{(\text{Old}\rightarrow\text{Young})}}{\text{Steps}_{(\text{Young}\rightarrow\text{Old})}}. \tag{2}$$

Young-associated:
teenager, millennial, young adult, Gen Z, college student, intern, junior, emerging, fresh, new graduate

Old-associated:
senior, elderly, retiree, boomer, veteran, experienced, mature, seasoned, elder, golden years

Figure 2: Age-associated terms used for generating answers.

4 Results

We conducted two experiments using the *Young* and *Old* datasets on each of the models. The first experiment is designed to demonstrate bias in LLMs and the second experiment which is conducted on Qwen 1.8B is designed to show that the results are reliable and to validate reproducibility.

Qwen 1.8B Results In the first experiment, we measured how many training steps were required and how long it took each model to reach 85% accuracy. The results in Figure 4a show that the model trained first on the young dataset, then on the old dataset ($M_{Y\rightarrow O}$) reached 85% accuracy the fastest, in about 50 training steps and 358 seconds. On the other hand, the model trained on the young dataset (M_Y) had the largest number of training steps and reached 85% accuracy in 100 training steps and 716 seconds. The other two models performed between these extremes: the model trained only on the old dataset (M_O) and the model trained first on the old dataset, then on the young dataset ($M_{O\rightarrow Y}$). The M_O model took about 75 training steps and 539 seconds, while the $M_{O\rightarrow Y}$ model took about 85 steps and 608 seconds. The total training time for each model in this experiment is presented in Table 1.

The second experiment was done with the same scheme but for a fixed 10 epochs and repeated five times with random seeds. The results of this experiment are shown in Figure 5, which plots the mean token accuracy and the confidence interval from these 5 experiments. These results show that the $M_{Y\rightarrow O}$ model converges fastest, consistent with the previous experiment, followed by the M_O, $M_{O\rightarrow Y}$, and finally M_Y. Additionally, similar to the first experiment, we can see that $M_{O\rightarrow Y}$ starts out with faster initial accuracy, but M_O eventually catches up and reaches a better final accuracy.

We also performed a statistical significance test on the 5 runs of the second experiment to ensure reproducibility. As shown in Figure 5, the narrow 98% confidence intervals demonstrate consistent results across runs for each model. Additionally, the non-overlapping intervals between different models confirm statistically significant performance differences between the four training conditions.

To understand if the observed differences in our results are statistically significant, we used two statistical tests: Analysis of Variance (ANOVA) [20] and Tukey's honestly significant difference

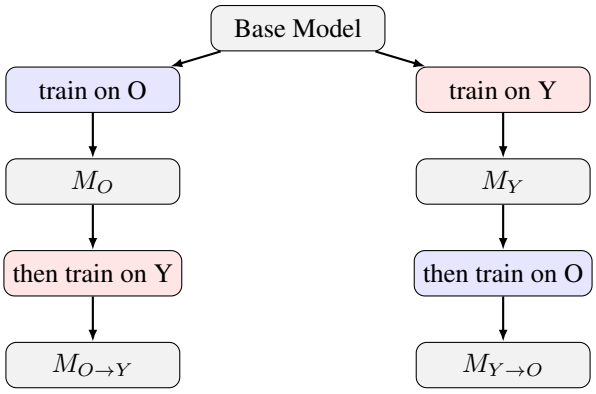
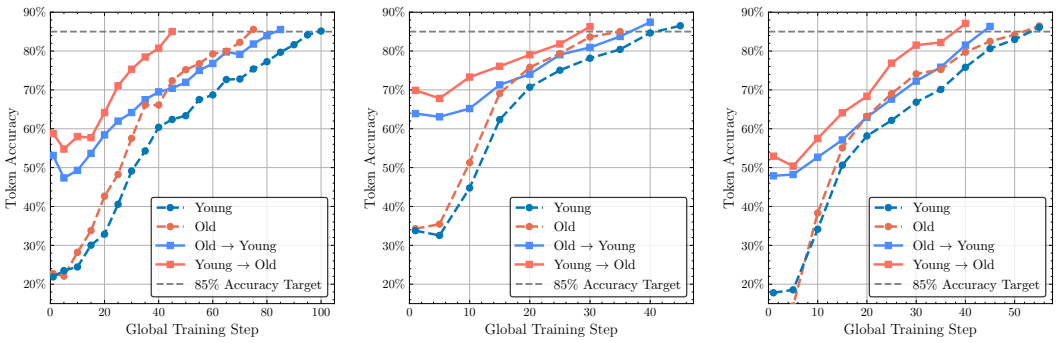


Figure 3: Summary of experiment one: The base model is trained on dataset Y (Young dataset) or dataset O (Old dataset), producing M_Y and M_O, which are then trained on the opposite dataset to yield $M_{Y\to O}$ and $M_{O\to Y}$.



(a) Qwen 1.8B (b) Llama 2 7B (c) Yi 6B

Figure 4: Experiment one with four configurations for Qwen, Llama 2, and Yi.

(HSD) test [21]. We first utilized a one-way ANOVA test to determine if any significant performance variation exists across the four training conditions. The null hypothesis (H0) for this test was that the mean final accuracies of all four groups are equal. The results showed an F-statistic of 2779.1 with a p-value < 0.001, which leads us to reject the null hypothesis. This confirms that a statistically significant difference exists somewhere among the four conditions. To identify which specific pairs of conditions differ, we then applied Tukey's HSD test. The results of this post-hoc analysis demonstrated that all six pairwise comparisons were statistically significant ($p < 0.05$), as every comparison rejected the null hypothesis of equal means. This provides strong evidence that the model performs distinctly in each of the four cases, supporting the interpretation that the model exhibits biased behavior toward age-associated language patterns.

Llama 7B Results Similar to the example of Qwen 1.8B, here we also run the first experiment with young and old datasets and report the results on the number of steps it takes for each setting to reach 85% accuracy. Results in Figure 4b show that the model trained first on the young dataset, then on the old dataset ($M_{Y\to O}$) reached 85% accuracy the fastest, in 30 training steps and 641 seconds. On the other hand, the model trained on the young dataset (M_Y) had the largest number of training steps and reached 85% accuracy in about 45 training steps and 1068 seconds. The other two models performed between these two. The M_O model took about 35 training steps and 746 seconds, while the $M_{O\to Y}$ model took 40 steps and 861 seconds to reach the 85% threshold.

Yi 6B Results The Yi model was trained with diverse and balanced instruction-tuning data, which may contribute to its more uniform performance across domains. The results show that unlike other two models, Yi 6B converges with the same number of steps (about 55 steps and 1400 seconds) on the young and old datasets, as shown in figure 4c. This indicates balanced treatment of both young M_Y

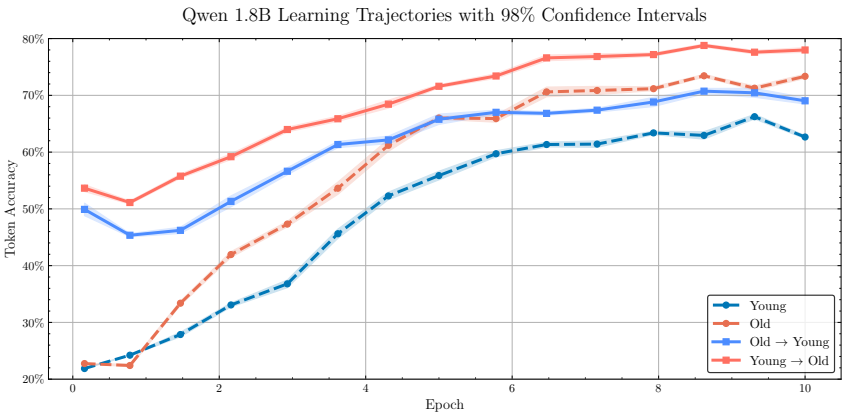


Figure 5: Results of Experiment two with four configurations across 5 random seeds, including mean performance and a confidence interval of 98%.

and old M_O datasets and minimal preference in sequential fine-tuning approaches (40 steps and 809 seconds in the young then old model and 45 steps and 1003 seconds in the old then young model).

5 Discussion

Table 1: Time to Reach (TTR) accuracy at 85% in seconds, and mean training speed (iter/s). The accuracy is defined as the number of correctly predicted tokens divided by the total number of tokens.

Model	Qwen 1.8B		Llama 2 7B		Yi 6B	
Dataset	TTR 85%	Speed	TTR 85%	Speed	TTR 85%	Speed
Old	539.0	0.139	746.0	0.047	1401.0	0.050
Young	716.0	0.139	1068.0	0.047	1403.0	0.050
Old → Young	608.0	0.140	861.0	0.046	1003.0	0.050
Young → Old	358.0	0.139	641.0	0.047	809.0	0.049

Our analysis reveals distinct bias patterns across architectures. Both Qwen 1.8B (33% Bias Score, 1.88 asymmetry) and Llama 7B (28% Bias Score, 1.33 asymmetry) demonstrate clear preference for "old" vocabulary, with faster convergence and lower training steps required. In contrast, Yi 6B shows balanced behavior (zero Bias Score, 1.12 asymmetry), suggesting similar treatment of both vocabulary sets. This variation across models validates our methodology's ability to detect both the presence and absence of bias, which makes it a reliable tool for model evaluation and selection. The results suggest that the Yi model exhibited less bias in our experiments compared to other models. Given that Yi uses a similar architecture to Llama 2, we hypothesize that this performance may arise from Yi's balanced instruction tuning and high-quality dataset. This has two implications: first, training data composition and training methods may play a stronger role than architecture in shaping bias. Second, our method could potentially be used to evaluate the data quality of different models in the future.

The consistency between Qwen and Llama, combined with Yi's balanced performance, demonstrates that our PEFT-based approach successfully captures genuine representational preferences rather than random variation or task difficulty differences. Understanding these bias patterns has important implications for reasoning reliability. Models exhibiting demographic biases may produce inconsistent conclusions when reasoning about identical scenarios presented with different age-associated vocabulary. In domains like healthcare or legal analysis, such inconsistencies could lead to systematically different recommendations based on implicit demographic assumptions rather than logical merit.

The modular nature of PEFT also suggests this approach could be extended to test other forms of bias that might similarly affect reasoning consistency. This could establish a comprehensive bias evaluation framework for critical applications.

6 Conclusion & Future Work

We proposed a new method to measure bias in LLMs using PEFT with two contrasting datasets and evaluated the method on three language models. This approach contributes to explainability in large models while remaining computationally efficient, since PEFT methods are designed to require fewer resources. Additionally, measuring bias from model outputs is challenging, often requiring human evaluation for precise conclusions.

We acknowledge that the scope of this work is limited to age-related vocabulary bias and a synthetic dataset that might not sufficiently represent the full scope of bias in language models. The proposed method can, however, become an initial stepping stone in the study of bias in large models through language proxies like the dataset types used here. This can serve as a basis for future work, which will potentially lead to a comprehensive method to study bias and the debiasing of models.

Studying different PEFT methods, exploring efficiency techniques [22], and normal fine-tuning, as well as using more diverse datasets including real-world opposing data, are some of the potential directions to explore. Future work could also examine how these bias detection methods relate to reasoning consistency, testing whether models with lower bias scores exhibit more reliable reasoning across demographic contexts. This approach can also be used to evaluate the training methodology of language models and potentially serve as a reliable tool in assessing and auditing evolving models such as those regularly updated in healthcare scenarios. We hope this work will help establish more explainable approaches, along with more balanced methods for training and evaluation, while keeping such approaches efficient and accessible.

References

[1] Aaron Chatterji, Thomas Cunningham, David J Deming, Zoe Hitzig, Christopher Ong, Carl Yan Shan, and Kevin Wadman. How people use chatgpt. Technical report, National Bureau of Economic Research, 2025.

[2] Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. Explainability for large language models: A survey. *ACM Transactions on Intelligent Systems and Technology*, 15(2):1–38, 2024.

[3] Roberto Navigli, Simone Conia, and Björn Ross. Biases in large language models: origins, inventory, and discussion. *ACM Journal of Data and Information Quality*, 15(2):1–21, 2023.

[4] Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. Bias and fairness in large language models: A survey, 2024.

[5] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

[6] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

[7] Linus Ta-Lun Huang and Tsung-Ren Huang. Generative bias: widespread, unexpected, and uninterpretable biases in generative models and their implications. *AI & SOCIETY*, pages 1–13, 2025.

[8] Afra Feyza Akyürek, Muhammed Yusuf Kocyigit, Sejin Paik, and Derry Wijaya. Challenges in measuring bias via open-ended language generation. *arXiv preprint arXiv:2205.11601*, 2022.

[9] Ewoenam Kwaku Tokpo and Toon Calders. Text Style Transfer for Bias Mitigation using Masked Language Modeling. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop*, pages 163–171, 2022.

[10] Paul Pu Liang, Chiyu Wu, Louis-Philippe Morency, and Ruslan Salakhutdinov. Towards understanding and mitigating social biases in language models. In *International conference on machine learning*, pages 6565–6576. PMLR, 2021.

[11] Haoyan Luo and Lucia Specia. From understanding to utilization: A survey on explainability for large language models. *arXiv preprint arXiv:2401.12874*, 2024.

[12] John Hewitt and Christopher D Manning. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138, 2019.

[13] Hao Peng, Xiaozhi Wang, Shengding Hu, Hailong Jin, Lei Hou, Juanzi Li, Zhiyuan Liu, and Qun Liu. Copen: Probing conceptual knowledge in pre-trained language models. *arXiv preprint arXiv:2211.04079*, 2022.

[14] Xu Wang, Yan Hu, Wenyu Du, Reynold Cheng, Benyou Wang, and Difan Zou. Towards understanding fine-tuning mechanisms of LLMs via circuit analysis. *arXiv preprint arXiv:2502.11812*, 2025.

[15] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. Qwen Technical Report. *arXiv preprint arXiv:2309.16609*, 2023.

[16] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[17] Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Guoyin Wang, Heng Li, Jiangcheng Zhu, Jianqun Chen, et al. Yi: Open foundation models by 01. ai. *arXiv preprint arXiv:2403.04652*, 2024.

[18] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

[19] Yuze Zhao, Jintao Huang, Jinghan Hu, Xingjun Wang, Yunlin Mao, Daoze Zhang, Zeyinzi Jiang, Zhikai Wu, Baole Ai, Ang Wang, Wenmeng Zhou, and Yingda Chen. SWIFT:A Scalable lightWeight Infrastructure for Fine-Tuning, 2024.

[20] Lars St, Svante Wold, et al. Analysis of variance (ANOVA). *Chemometrics and intelligent laboratory systems*, 6(4):259–272, 1989.

[21] Hervé Abdi, Lynne J Williams, et al. Tukey's honestly significant difference (HSD) test. *Encyclopedia of research design*, 3(1):1–5, 2010.

[22] Pouya Shaeri and Ariane Middel. Mid-l: Matrix-interpolated dropout layer with layer-wise neuron selection. *arXiv preprint arXiv:2505.11416*, 2025.