Interspeech 2024
1-5 September 2024, Kos, Greece

Non-Linear Inference Time Intervention: Improving LLM Truthfulness

Jakub Hoscilowicz[*1,3], *Adam Wiacek*[*1], *Jan Chojnacki*[1,2], *Adam Cieslak*[1], *Leszek Michon*[1], *Artur Janicki*[3]

[1]Samsung R&D Institute Poland, Warsaw, Poland
[2]University of Warsaw, Poland
[3]Warsaw University of Technology, Poland

{j.hoscilowic, a.wiacek2, a.cieslak, l.michon}@samsung.com,
jr.chojnacki@uw.edu.pl, artur.janicki@pw.edu.pl

Abstract

In this work, we explore LLM's internal representation space to identify attention heads that contain the most truthful and accurate information. We further developed the Inference Time Intervention (ITI) framework, which lets bias LLM without the need for fine-tuning. The improvement manifests in introducing a non-linear multi-token probing and multi-token intervention: Non-Linear ITI (NL-ITI), which significantly enhances performance on evaluation benchmarks. NL-ITI is tested on diverse multiple-choice datasets, including TruthfulQA, on which we report over $14\,\%$ relative MC1 (accuracy of model pointing to the correct answer) improvement with respect to the baseline ITI results. Moreover, we achieved a 10% relative improvement over the recently released Truth Forest (TrFf) method that also focused on ITI improvement.

Index Terms: Large Language Models, Representation Editing, Probing, AI Ethics, AI Safety

1. Introduction

Large Language Models (LLMs) are a major achievement in the domain of artificial intelligence, particularly within natural language processing (NLP). Their capabilities span a wide array of applications, from generating human-like texts to understanding and processing complex language structures. However, the probabilistic nature of these models often gives rise to certain challenges, including the phenomena of hallucinations [1, 2] and the generation of toxic content [3]. LLM models trained on extensive datasets inadvertently absorb and repeat cultural, gender-based, racial, or ideological biases in their training dataset [4, 5]. These issues underscore the motivation behind the development of robust benchmarks and methodologies aimed at evaluating and enhancing the safety, fairness, and accuracy of LLM outputs. A comprehensive strategy is necessary that includes diversifying training datasets, developing algorithms to detect and neutralize bias, and implementing robust testing protocols for biased outputs. Recent developments suggest how to assess and mitigate the model's bias [6, 7, 8, 9, 10].

Our work investigates the internal representation space of LLMs to identify and utilize the most informative attention heads for specific tasks. During inference, the activations of such heads are modified, thus refining LLM-generated content. Our primary contribution is a notable enhancement of the Inference Time Intervention (ITI) method [7], leading to higher performance on LLM benchmarks and better generalization capability. The improvements manifest in two distinct aspects: firstly, the introduction of non-linearity to the probing model, which facilitates a more effective identification of attention heads collecting the type of desired knowledge (e.g., truthfulness). Secondly, the employment of an expanded token context during interventions enables a more refined construction of the intervention vector, thereby directing attention heads more effectively toward truthfulness. This enhanced construction of the intervention vector is attributed to the observation that truthful knowledge is not solely concentrated in the vector corresponding to the final token, but is distributed across a broader context. We discuss how our framework can be used to bias LLM toward any abstract concept (truthfulness, correctness, toxicity-prevention). We present our advancements and their contribution to developing safer, more accurate, and ethically responsible LLM systems, demonstrating the potential of our approach for future AI applications.

[*]The first two authors made equal contribution.

2. Related Work

Efforts to mitigate LLM biases have led to the development of diverse strategies, of which one of the most impactful is the Reinforcement Learning from Human Feedback (RLHF) [11]. It aligns models with human feedback, reducing bias by adjusting model behaviors based on human preferences. Such an approach, while effective, demands significant human labor [12], highlighting the need for novel, automated bias mitigation methods.

In contrast, approaches like ITI [7] act more directly by modifying the model's internal representations. It was noticed that the LLMs sometimes 'know' they produce false statements [13]. This motivated the ITI authors to bias the model towards more truthful behavior. This involves a two-step process, where attention heads are first evaluated for their accuracy, and then Mass Mean Shift vectors are applied to a subset of top-performing heads during inference. These vectors, calculated as the mean difference between activations for true and false answers, are pre-computed and attached to the model, enabling a precision improvement with minimal computational overhead.

Recently, another work focusing on analyzing and improving the probing procedure of the internal representation space has been reported. Authors of [14] introduced the Truth Forest (TrFr) employing multi-dimensional orthogonal probes. However still, they focused on optimizing the original ITI framework.

The increasing demand for LLMs in a practical text generation underscores the importance of model fairness and truthfulness. Benchmarks such as TruthfulQA [10] have been introduced to evaluate model truthfulness across various domains, including health, law, finance, and politics. These benchmarks

10.21437/Interspeech.2024-819

challenge models with questions designed to elicit imitative falsehoods, thereby testing the model's ability to maintain truthfulness across topics. Similarly, datasets like BBQ [9] and ToxiGen [8] assess LLM fairness and ability to handle nuanced manifestations of hate speech.

Recent work has sparked an interest in LLM personality categorization and psychometrics. Contributions in this area include adjusting LLM personality traits [15], using the Myers-Briggs Type Indicator (MBTI) for evaluation [16], and simulating diverse personalities in models [17]. Our work builds upon these foundational efforts, aiming to contribute to the development of more ethical and unbiased LLM applications.

3. Method

As reported in [7], LLMs seem to preserve an internal representation of abstract concepts such as truth and honesty, even though they do not generate factual responses. Simple prompt stimulation may not be enough to access the full potential stored in the internal representations. In [7], one uses labeled data (i.e. question-answer Q+A pairs from TruthfulQA split) to train a linear probing model, which identifies attention heads storing the truthful representations. For each such head, the truthful direction is calculated. During the inference, attention head activations are shifted in the truthful direction.

This method unfolds in two phases. Initially, a linear probing model is trained on representations returned by the attention heads for a given probing trainset. The assumption is that the higher the accuracy of the probing model, the higher the amount of desired knowledge (e.g., truthful). Mathematically, the probing operation p_θ may be described as:

$$p_\theta\left(x_l^h\right) = \text{sigmoid}\left(\left\langle \theta, x_l^h \right\rangle\right), \tag{1}$$

where θ is a set of trainable parameters, and x_l^h is an activation of token x at head h and layer l. In the process of probe training, N question-answers pairs are concatenated and the activations that correspond to the last token $\left\{(x_l^h, y)_i\right\}_{i=1}^N$ are collected (where y is a binary label that indicates true or false answer).

Note that the probes need to be trained on a labeled dataset reflecting the concept (here, we used TruthfulQA [10]). For each concept (e.g., truthfulness, toxicity-prevention, personality adjustment), a different biasing dataset and probe training are necessary.

Following [7], the intervention is then given by:

$$x_{l+1} = x_l + \sum_{h=1}^{H} Q_l^h \left(\text{Att}_l^h\left(P_l^h x_l\right) + \alpha \sigma_l^h \theta_l^h\right), \tag{2}$$

where $x_l \in \mathbb{R}^{DH}$ is the l-th token high-dimensional embedding, $P_l^h \in \mathbb{R}^{D\times DH}$ is a mapping operator from token embedding to D-dimensional attention head space, and Att_l^h operator connects information from other tokens and gives us activations seen in Equation 1: $x_l^h = \text{Att}_l^h\left(P_l^h x_l\right)$. Finally, the last term in Equation 2 is the intervention term. It can be understood as follows: when calculating the next-token-prediction x_{l+1} the residual stream (previous x_l token and the weighted sum of activations) is modified by adding the biasing direction θ_l^h multiplied by its standard deviation (with respect to all of the Q+A pairs) and the *intervention strength* α.

To calculate the biasing direction for each head and layer, one takes an average of activations of the last token among the Q+A dataset:

$$\theta_l^h = \frac{1}{N}\sum_{i=1}^{N}\left(x_l^h\right)_i, \quad h \in \text{Top Heads}. \tag{3}$$

The biasing directions are appended only for a number of attention heads with the highest probing accuracy. Their number is controlled by a second hyperparameter K.

4. Proposed improvement

ITI [7] uses attention head probing based on a logistic regression probing model, as shown in Equation 1. We think that it does not optimally capture the complexity of the concept representation in the activation space.

Therefore, we propose improved probing and suggest using non-linear MultiLayer Perceptron (MLP) as the probe, changing Equation 1 to:

$$p_\theta\left(x_l^h\right) = \text{MLP}\left(\left\langle \theta, x_l^h \right\rangle\right). \tag{4}$$

Improving probes' accuracy leads to a more appropriate choice of the top heads. The top heads are then used in the ITI procedure described in Equation 2. Non-linear probing has generally higher information capacity and is able to capture more of the inherent linguistic information in the representation [18, 19, 20].

Moreover, instead of using only the last token for probing training, we focus on the average optimal number of last tokens. Increasing the information capacity of the probing model can be naturally followed by providing more information encoded in multiple tokens provided to the MLP.

This modification can be mathematically expressed as a change in which one collects the training dataset:

$$\left\{\left(x_l^h, y\right)_i\right\}_{i=1}^N \xrightarrow{\text{multi-token}} \left\{\left(\left\langle x_l^h \right\rangle_\tau, y\right)_i\right\}_{i=1}^N, \tag{5}$$

where instead of taking just the last token, on which the activation x_l^h are calculated, we take an average of τ last tokens. This optimum is found experimentally and described in detail in Section 5.

Similarly, extending the number of tokens used in intervention is also relevant. This is our second improvement to the framework.

In particular, during the inference, at each attention head, we add biasing directions corresponding to the mean of the last ρ tokens in a Q+A pair. To calculate the biasing direction vector we average over all Q+A pairs as in Equation 3:

$$\theta_l^h = \frac{1}{N}\sum_{i=1}^{N}\left(\left\langle x_l^h \right\rangle_\rho\right)_i, \quad h \in \text{Top Heads}. \tag{6}$$

As before, the number of tokens ρ, over which the activations are averaged, needs to be found empirically.

From now on, we will call this framework *Non-Linear-Inference Time Intervention* (NL-ITI).

5. Experiments

We verify how token addition and probing non-linearity affect evaluation metrics. We discuss the generalization capabilities of NL-ITI and show the performance improvement on diverse reasoning tests.

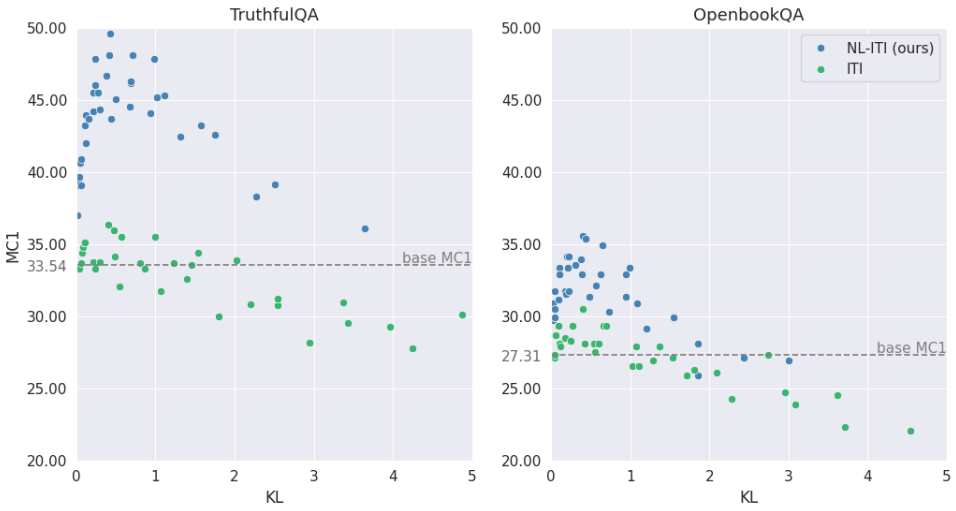

Figure 1: *How MC1 correlates with KL divergence. The results were collected for ITI and NL-ITI using different hyperparameter sets (α, heads to intervene) for TruthfulQA and OpenBookQA datasets. On each benchmark, baseline LLaMA-2-7B performance is shown.*

5.1. Evaluation

Throughout this work, all experiments were performed on LLaMA-2-7B [21] model. This open-source LLM has been used extensively in previous work on AI safety. Particularly, it let us to directly compare NL-ITI to related methods [7, 14].

Figure 1 shows how ITI and NL-ITI compare on two benchmarks. Each point corresponds to a different hyperparameter set, reflected in different Kullback-Leibler divergence (KL) calculated with respect to OpenWebText [22] distribution. At each test-point, NL-ITI outperforms ITI. NL-ITI creates a peak in KL ~ 0.5, the correlation is also non-linear around this region.

In [7], the authors report MC1 and MC2 scores introduced in TruthfulQA [10] benchmark and generative evaluation methods. MC1 and MC2 can be understood as the accuracy at which the model predicts the correct answer for a given question, if the model output was to be restricted only to generate one of the (correct or otherwise) answers specified in the dataset. MC1 is applicable for only single-correct answer datasets, while MC2 is well-defined also for multiple-correct datasets. For the implementation of MC1 and MC2 scores we reference [10] code and our GitHub repository[1]. We proceed with MC1 and MC2 score evaluation, as they are more reliable and replicable independently of the Judge LLM Models. MC-scores do not depend on the underlying judging model reasoning capabilities, since labeled data is used in the accuracy calculation. Moreover, this allows us for more direct comparisons with other approaches [7, 14], as the evaluations with closed-source GPT-4 may depend on the OpenAI software updates.

Similarly to [7], Cross Entropy (CE) and KL divergence are calculated to see how much the intervention result diverges from the original LLaMA-2-7B [21] token distribution. For both metrics, lower values correspond to less change in the model's behavior. Very large CE and KL values suggest intervention procedure changed the output token distribution in a major way. One could imagine that such a change could negatively affect LLM's language comprehension. Hence, these values are treated as a *sanity check* and may suggest that the generalization capabilities of the model are impaired at very large CE and KL values. However, it is difficult to estimate at which values these metrics point to the generalization collapse. Moreover, slightly more divergent next-token prediction distribution could still lead to better-performing LLM. We report that NL-ITI gives a larger MC1 than ITI for every value of KL (see Figure 1).

Based on the results shown in Table 2 (major generalization test) we can see that NL-ITI outperforms ITI on ARC [23], MMLU [24], and OpenBook [25] benchmarks with a slightly higher CE ($\sim$10 %). Therefore, we suggest that KL and CE should be treated as a sanity check, while the generalization capabilities should be evaluated on diverse benchmarks. To better visualize our point, we provide two plots in Figure 1 with function MC1 of KL on two datasets: TruthfulQA and OpenBookQA. It is clear that the initial assumption of correlation between MC1 and KL made in [7] was true. However, for NL-ITI, there exists a local maximum of MC1 away from KL $= 0$.

5.2. Results

Combining the three simple adjustments described in Section 4 leads to surprising performance improvements over the original ITI approach. As can be seen in Table 1, the MC1 score has improved by 50 % relative to the baseline LLM result with no intervention. The effect of the intervention is 13 % relative higher than with the baseline ITI[2]. With enhancements described in the previous section, we have managed to vastly improve the model truthfulness, compared to the baseline ITI approach.

As described in Section 4, increasing the capacity of probing through a higher amount of hidden neurons and non-linear activations leads to better estimation of knowledge amount in the attention heads. We tested the group of MLP models, with a growing number of parameters, against the logistic regression model. We found that adding a middle layer between the input layer and the sigmoid layer gave the best results. The more advanced models performed worse, probably because they overfit to the probing training data, which is limited in the case of TruthfulQA ($\sim$800 samples). Interestingly, the effect of apply-

[1] https://github.com/Samsung/NL-ITI

[2] For direct comparison, we recalculated this score, replicating the original methodology [7], achieving MC1 = 36.35 %.

Table 1: *Comparison between baseline (LLaMA-2-chat-7B model, TruthfulQA dataset), ITI and NL-ITI. Compared values have all been achieved with few-shot-prompting.*

Model	MC1 [%]	MC2 [%]	CE	KL
LLaMA-2-7B	33.54	50.34	2.53	0.00
ITI	36.35	54.72	2.65	0.40
TrFr	39.30	-	2.59	0.22
NL-ITI (ours)	**50.19**	**67.73**	2.85	0.43
w/o optimized probe	42.96	61.48	2.66	0.25
w/o multi-token	40.75	59.83	3.33	1.40

Table 2: *Comparison of generalization of ITI and NL-ITI on out-of-distributions benchmarks: AI2's Reasoning Challenge, Massive Multitask Language Understanding, and OpenBookQA*

Model	Dataset	MC1 [%]	MC2 [%]	CE	KL
LLaMA-2-7B	ARC	41.20	40.69	2.53	0.00
ITI	ARC	40.34	38.78	2.54	0.12
NL-ITI (ours)	ARC	**44.27**	**43.20**	2.82	0.40
LLaMA-2-7B	MMLU	38.48	38.66	2.53	0.00
ITI	MMLU	38.55	38.27	2.58	0.04
NL-ITI (ours)	MMLU	**40.31**	**39.82**	2.60	0.10
LLaMA-2-7B	OBQA	27.31	26.36	2.53	0.00
ITI	OBQA	30.52	28.26	2.82	0.40
NL-ITI (ours)	OBQA	**33.94**	**32.65**	2.86	0.32

ing non-linear probing to the attention heads was the most profound in the first six layers (Figure 2). Moreover, the non-linear probing points to the fact that truthful knowledge is much more diffused along the attention heads than linear probing would suggest. The matrix in Figure 3 summarizes our results. The rows correspond to the number of tokens used during probe training, specifically to the number τ defined in Equation 5. Columns present the influence of increasing the token number used in the intervention; this corresponds to ρ in Equation 6. The reported impact of using an increased number of tokens is particularly strong in probing. However, having the increased number of tokens both in probing and intervention produces a joint effect and yields the optimum at $(\rho, \tau) = (6, 4)$.

Our results suggest that a significant amount of information about the concept (truthfulness) might be contained not only in the vector corresponding to the last token of LLM's answer, but also in preceding vectors. For reference, in the ITI approach, only the activations from the last token were used in probing and intervention.

6. Conclusions

In this work, we proposed modifications that significantly improved the accuracy of the ITI method (as measured by major public benchmarks). Our optimizations included the application of an MLP during the probing procedure, which increased the precision of finding attention heads with the best internal representation of the desired (truthful) type of knowledge. Similarly, going beyond the last token activations during representation engineering, inference, and ultimately intervention, led the model generation to the desired outcome (i.e., truthfulness).

Our optimized model NL-ITI outperformed other techniques on four major benchmarks, including TruthfulQA, on

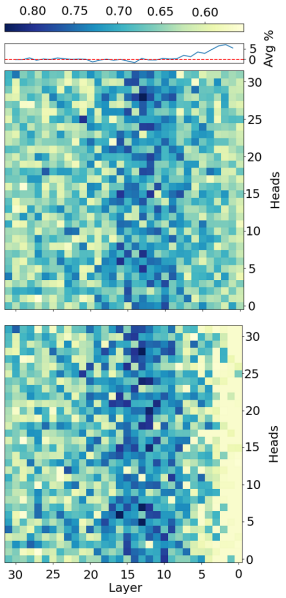

Figure 2: *Probing accuracy for each attention head of the LLM on TruthfulQA dataset for linear probing (ITI) – bottom, and non-linear probing (NL-ITI) – top. Accuracy results are 'smoothed' between neighboring attention heads (lower standard deviation).*

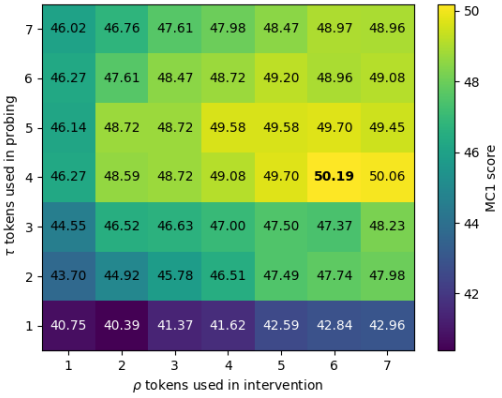

Figure 3: *Heat map of MC1 evaluation scores of TruthfulQA dataset for different combinations of number of tokens used during probing and intervention. The best performing model corresponds to* $(\rho, \tau) = (6, 4)$.

which we report over $14\,\%$ relative MC1 metric improvement (up to $50.19\,\%$) with respect to the baseline ITI results ($36.35\,\%$). Additionally, we achieved significant improvement ($+10\,\%$ relative) over the recently released TrFR method that also focused on ITI improvement. The need for labeled data is a limiting factor of NL-ITI approach, and future research could explore use of unsupervised models. Guiding LLM's internal representations is a promising direction for ensuring safe, truthful, and more human-centric AI. Such an approach is more data-efficient than fine-tuning (see fine-tuning efficiency discussion [26]) and more labor-efficient than human reinforcement (see other approaches, e.g., in [12]).

7. References

[1] P. Manakul, A. Liusie, and M. Gales, "SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, H. Bouamor, J. Pino, and K. Bali, Eds. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 9004–9017. [Online]. Available: https://aclanthology.org/2023.emnlp-main.557

[2] J. Li, X. Cheng, X. Zhao, J.-Y. Nie, and J.-R. Wen, "HaluEval: A large-scale hallucination evaluation benchmark for large language models," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, H. Bouamor, J. Pino, and K. Bali, Eds. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 6449–6464. [Online]. Available: https://aclanthology.org/2023.emnlp-main.397

[3] O. Shaikh, H. Zhang, W. Held, M. Bernstein, and D. Yang, "On second thought, let's not think step by step! bias and toxicity in zero-shot reasoning," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, A. Rogers, J. Boyd-Graber, and N. Okazaki, Eds. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 4454–4470. [Online]. Available: https://aclanthology.org/2023.acl-long.244

[4] A. Taubenfeld, Y. Dover, R. Reichart, and A. Goldstein, "Systematic biases in LLM simulations of debates," *arXiv Preprint, arXiv:2402.04049*, 2024.

[5] K.-C. Yeh, J.-A. Chi, D.-C. Lian, and S.-K. Hsieh, "Evaluating interfaced LLM bias," in *Proceedings of the 35th Conference on Computational Linguistics and Speech Processing (ROCLING 2023)*, 2023, pp. 292–299.

[6] A. Zou, L. Phan, S. Chen, J. Campbell, P. Guo, R. Ren, A. Pan, X. Yin, M. Mazeika, A.-K. Dombrowski, S. Goel, N. Li, M. J. Byun, Z. Wang, A. Mallen, S. Basart, S. Koyejo, D. Song, M. Fredrikson, J. Z. Kolter, and D. Hendrycks, "Representation engineering: A top-down approach to ai transparency," 2023.

[7] K. Li, O. Patel, F. Viégas, H. Pfister, and M. Wattenberg, "Inference-time intervention: Eliciting truthful answers from a language model," in *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., vol. 36. Curran Associates, Inc., 2023, pp. 41 451–41 530. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2023/file/81b8390039b7302c909cb769f8b6cd93-Paper-Conference.pdf

[8] T. Hartvigsen, S. Gabriel, H. Palangi, M. Sap, D. Ray, and E. Kamar, "ToxiGen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, S. Muresan, P. Nakov, and A. Villavicencio, Eds. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 3309–3326. [Online]. Available: https://aclanthology.org/2022.acl-long.234

[9] A. Parrish, A. Chen, N. Nangia, V. Padmakumar, J. Phang, J. Thompson, P. M. Htut, and S. Bowman, "BBQ: A hand-built bias benchmark for question answering," in *Findings of the Association for Computational Linguistics: ACL 2022*, S. Muresan, P. Nakov, and A. Villavicencio, Eds. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 2086–2105. [Online]. Available: https://aclanthology.org/2022.findings-acl.165

[10] S. Lin, J. Hilton, and O. Evans, "TruthfulQA: Measuring how models mimic human falsehoods," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, S. Muresan, P. Nakov, and A. Villavicencio, Eds. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 3214–3252. [Online]. Available: https://aclanthology.org/2022.acl-long.229

[11] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. Christiano, J. Leike, and R. Lowe, "Training language models to follow instructions with human feedback," *arXiv Preprint, arXiv:2203.02155*, 2022.

[12] H. Lee, S. Phatale, H. Mansoor, T. Mesnard, J. Ferret, K. Lu, C. Bishop, E. Hall, V. Carbune, A. Rastogi, and S. Prakash, "RLAIF: Scaling reinforcement learning from human feedback with AI feedback," *arXiv Preprint, arXiv:2309.00267*, 2023.

[13] S. Kadavath, T. Conerly, A. Askell, T. Henighan, D. Drain, E. Perez, N. Schiefer, Z. Hatfield-Dodds, N. DasSarma, E. Tran-Johnson, S. Johnston, S. El-Showk, A. Jones, N. Elhage, T. Hume, A. Chen, Y. Bai, S. Bowman, S. Fort, D. Ganguli, D. Hernandez, J. Jacobson, J. Kernion, S. Kravec, L. Lovitt, K. Ndousse, C. Olsson, S. Ringer, D. Amodei, T. Brown, J. Clark, N. Joseph, B. Mann, S. McCandlish, C. Olah, and J. Kaplan, "Language models (mostly) know what they know," 2022.

[14] Z. Chen, X. Sun, X. Jiao, F. Lian, Z. Kang, D. Wang, and C.-Z. Xu, "Truth forest: Toward multi-scale truthfulness in large language models through intervention without tuning," *arXiv Preprint, arXiv:2312.17484*, 2024.

[15] S. Mao, N. Zhang, X. Wang, M. Wang, Y. Yao, Y. Jiang, P. Xie, F. Huang, and H. Chen, "Editing personality for LLMs," *arXiv Preprint, arXiv:2310.02168*, 2023.

[16] K. Pan and Y. Zeng, "Do LLMs possess a personality? Making the MBTI test an amazing evaluation for large language models," *arXiv Preprint, arXiv:2307.16180*, 2023.

[17] J. tse Huang, W. Wang, M. H. Lam, E. J. Li, W. Jiao, and M. R. Lyu, "Revisiting the reliability of psychological scales on large language models," *arXiv Preprint, arXiv:2305.19926*, 2023.

[18] T. Pimentel, J. Valvoda, R. H. Maudslay, R. Zmigrod, A. Williams, and R. Cotterell, "Information-theoretic probing for linguistic structure," *arXiv Preprint, arXiv:2004.03061*, 2020.

[19] J. C. White, T. Pimentel, N. Saphra, and R. Cotterell, "A non-linear structural probe," *arXiv Preprint, arXiv:2105.10185*, 2021.

[20] J. Hościłowicz, M. Sowański, P. Czubowski, and A. Janicki, "Can we use probing to better understand fine-tuning and knowledge distillation of the BERT NLU?" in *Proceedings of the 15th International Conference on Agents and Artificial Intelligence (ICAART), Volume 3, Lisbon, Portugal, February 22-24, 2023*, A. P. Rocha, L. Steels, and H. J. van den Herik, Eds. SCITEPRESS, 2023, pp. 625–632. [Online]. Available: https://doi.org/10.5220/0011724900003393

[21] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample, "Llama: Open and efficient foundation language models," 2023.

[22] A. Gokaslan, V. Cohen, E. Pavlick, and S. Tellex, "Openwebtext corpus," http://Skylion007.github.io/OpenWebTextCorpus, 2019.

[23] P. Clark, I. Cowhey, O. Etzioni, T. Khot, A. Sabharwal, C. Schoenick, and O. Tafjord, "Think you have solved question answering? try arc, the ai2 reasoning challenge," *arXiv:1803.05457v1*, 2018.

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

[25] T. Mihaylov, P. Clark, T. Khot, and A. Sabharwal, "Can a suit of armor conduct electricity? a new dataset for open book question answering," in *EMNLP*, 2018.

[26] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. de Laroussilhe, A. Gesmundo, M. Attariyan, and S. Gelly, "Parameter-efficient transfer learning for nlp," *arXiv Preprint, arXiv:1902.00751*, 2019.