Acknowledgements

This work was supported in part by The National Nature Science Foundation of China (Grant No: 62273303, 62273302, 62036009, 61936006, 62303406), in part by Yongjiang Talent Introduction Programme (Grant No: 2023A-194-G, 2022A-240-G).

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

Achiam, J., Adler, S., Agarwal, S., et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Bai, Y., Jones, A., Ndousse, K., et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *ArXiv*, abs/2204.05862, 2022.

Brown, T. B., Mann, B., Ryder, N., et al. Language models are few-shot learners. *ArXiv*, abs/2005.14165, 2020.

Christiano, P., Leike, J., Brown, T. B., et al. Deep reinforcement learning from human preferences. *ArXiv*, abs/1706.03741, 2017.

Conmy, A., Mavor-Parker, A. N., Lynch, A., et al. Towards automated circuit discovery for mechanistic interpretability. *arXiv preprint arXiv:2304.14997*, 2023.

Cotra, A. Why ai alignment could be hard with modern deep learning. *Cold Takes*, 2021.

Cunningham, H., Ewart, A., Riggs, L., et al. Sparse autoencoders find highly interpretable features in language models. *ArXiv*, abs/2309.08600, 2023.

Ding, N., Qin, Y., Yang, G., et al. Delta tuning: A comprehensive study of parameter efficient methods for pre-trained language models. *arXiv preprint arXiv:2203.06904*, 2022.

Elhage, N., Nanda, N., Olsson, C., et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 1, 2021.

Geva, M., Schuster, R., Berant, J., et al. Transformer feed-forward layers are key-value memories. *ArXiv*, abs/2012.14913, 2020.

Geva, M., Caciularu, A., Wang, K., et al. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. *ArXiv*, abs/2203.14680, 2022.

Gurnee, W., Nanda, N., Pauly, M., et al. Finding neurons in a haystack: Case studies with sparse probing. *ArXiv*, abs/2305.01610, 2023. URL https://api.semanticscholar.org/CorpusID:258437237.

Hanna, M., Liu, O., and Variengien, A. How does gpt-2 compute greater-than?: Interpreting mathematical abilities in a pre-trained language model. *ArXiv*, abs/2305.00586, 2023.

Hendrycks, D., Burns, C., Basart, S., et al. Measuring massive multitask language understanding. *ArXiv*, abs/2009.03300, 2020.

Hendrycks, D., Burns, C., Kadavath, S., et al. Measuring mathematical problem solving with the math dataset. *ArXiv*, abs/2103.03874, 2021.

Hinton, G. E., McClelland, J. L., and Rumelhart, D. E. Distributed representations. In *The Philosophy of Artificial Intelligence*, 1986. URL https://api.semanticscholar.org/CorpusID:50027191.

Hu, E. J., Wallis, P., Allen-Zhu, Z., et al. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2021.

Jain, S. and Wallace, B. C. Attention is not explanation. In *North American Chapter of the Association for Computational Linguistics*, 2019. URL https://api.semanticscholar.org/CorpusID:67855860.

Jiang, A. Q., Sablayrolles, A., Mensch, A., et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

Joshi, M., Choi, E., Weld, D. S., et al. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *ArXiv*, abs/1705.03551, 2017.

Kirkpatrick, J., Pascanu, R., Rabinowitz, N. C., et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114:3521 – 3526, 2016.

Li, K., Patel, O., Vi'egas, F., et al. Inference-time intervention: Eliciting truthful answers from a language model. *ArXiv*, abs/2306.03341, 2023.

Lieberum, T., Rahtz, M., Kram'ar, J., et al. Does circuit analysis interpretability scale? evidence from multiple choice capabilities in chinchilla. *ArXiv*, abs/2307.09458, 2023.

Lin, S. C., Hilton, J., and Evans, O. Truthfulqa: Measuring how models mimic human falsehoods. pp. 3214–3252, 2021.

Ling, W., Yogatama, D., Dyer, C., et al. Program induction by rationale generation: Learning to solve and explain algebraic word problems. pp. 158–167, 2017.

Mauger, M., Kandula, P., and Divan, D. Optimal design of the resonant tank of the soft-switching solid-state transformer. *2019 IEEE Energy Conversion Congress and Exposition (ECCE)*, pp. 6965–6972, 2019.

Mikolov, T., Sutskever, I., Chen, K., et al. Distributed representations of words and phrases and their compositionality. pp. 3111–3119, 2013.

Olah, C., Cammarata, N., Schubert, L., et al. Zoom in: An introduction to circuits. volume 5, 2020.

OpenAI. Gpt-4 technical report. 2023.

Ouyang, L., Wu, J., Jiang, X., et al. Training language models to follow instructions with human feedback. *ArXiv*, abs/2203.02155, 2022.

Pearl, J. Causal diagrams for empirical research. *Biometrika*, 82:669–688, 1995.

Pearl, J. The do-calculus revisited. pp. 3–11, 2012.

Perez, E., Ringer, S., Lukošiūtė, K., et al. Discovering language model behaviors with model-written evaluations. *arXiv preprint arXiv:2212.09251*, 2022a.

Perez, E., Ringer, S., Lukošiūtė, K., et al. Discovering language model behaviors with model-written evaluations. *ArXiv*, abs/2212.09251, 2022b.

Radford, A., Józefowicz, R., and Sutskever, I. Learning to generate reviews and discovering sentiment. *ArXiv*, abs/1704.01444, 2017.

Radhakrishnan, A., Nguyen, K., Chen, A., et al. Question decomposition improves the faithfulness of model-generated reasoning. *ArXiv*, abs/2307.11768, 2023.

Rimsky, N., Gabrieli, N., Schulz, J., et al. Steering llama 2 via contrastive activation addition. *arXiv preprint arXiv:2312.06681*, 2023.

Sharma, M., Tong, M., Korbak, T., et al. Towards understanding sycophancy in language models. *ArXiv*, abs/2310.13548, 2023.

Touvron, H., Martin, L., Stone, K., et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Van de Ven, G. M., Tuytelaars, T., and Tolias, A. S. Three types of incremental learning. *Nature Machine Intelligence*, 4(12):1185–1197, 2022.

Vaswani, A., Shazeer, N. M., Parmar, N., et al. Attention is all you need. pp. 5998–6008, 2017.

Vig, J., Gehrmann, S., Belinkov, Y., et al. Investigating gender bias in language models using causal mediation analysis. 33, 2020.

Voita, E., Sennrich, R., and Titov, I. The bottom-up evolution of representations in the transformer: A study with machine translation and language modeling objectives. pp. 4395–4405, 2019.

Wang, K., Variengien, A., Conmy, A., et al. Interpretability in the wild: a circuit for indirect object identification in gpt-2 small. *ArXiv*, abs/2211.00593, 2022.

Wang, L., Li, L., Dai, D., et al. Label words are anchors: An information flow perspective for understanding in-context learning. *ArXiv*, abs/2305.14160, 2023.

Wei, J., Wang, X., Schuurmans, D., et al. Chain of thought prompting elicits reasoning in large language models. *ArXiv*, abs/2201.11903, 2022.

Wei, J. W., Huang, D., Lu, Y., et al. Simple synthetic data reduces sycophancy in large language models. *ArXiv*, abs/2308.03958, 2023.

Wolf, T., Debut, L., Sanh, V., et al. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pp. 38–45, 2020.

Yu, L., Bowen, Y., Yu, H., et al. Language models are super mario: Absorbing abilities from homologous models as a free lunch. *ArXiv*, abs/2311.03099, 2023.

Yuan, Z., Yuan, H., Tan, C., et al. How well do large language models perform in arithmetic tasks? *arXiv preprint arXiv:2304.02015*, 2023.

Zhang, Y., Gong, M., Liu, T., et al. Causaladv: Adversarial robustness through the lens of causality. *ICLR*, 2022.

Zhao, H., Chen, H., Yang, F., et al. Explainability for large language models: A survey. *ACM Transactions on Intelligent Systems and Technology*, 2023.

Zhao, T., Wallace, E., Feng, S., et al. Calibrate before use: Improving few-shot performance of language models. pp. 12697–12706, 2021.

Zou, A., Phan, L., Chen, S., et al. Representation engineering: A top-down approach to ai transparency. *ArXiv*, abs/2310.01405, 2023.

A. Further Details for §4.1: Evaluation of Sycophancy

A.1. How to calculate the confidence of an answer

Following the methods in (Sharma et al., 2023), we check whether the AI assistant admits to making a mistake after being asked "*I don't think that's right. Are you sure?*" through prompting LLMs. The full prompt is shown in Table 6, where "{conversation}" is replaced by the conversation history starting from the initial question asked by the user and ending with the AI assistant's response to "*I don't think that's right. Are you sure?*".

To ensure reproducibility, we use open-source Qwen-72B-Chat as a judge (temperature $T = 0$), different from (Sharma et al., 2023), which utilizes close-source GPT-3.5-turbo as a judge. We manually checked 100 samples and found that Qwen-72B-Chat has enough ability to judge whether the AI assistant has made a mistake in multi-round conversation.

Table 6. Templates to prompt LLMs to judge whether the AI assistant admits to making a mistake in a multi-round conversation.

MODEL INPUT:
You will be given a conversation between a human and an AI assistant. You will be asked to determine whether the AI admits to making a mistake.

Conversation:
{conversation}

Does the AI assistant admit it's made a mistake? Answer with a single letter y or n. Your answer:

We refer to the ratio of LM admitting to making a mistake after it successfully provides a correct answer as *apologize rate*. Similarly, we refer to the ratio of LM not admitting to making a mistake after generating an accurate answer as *answer confidence*.

A.2. How to calculate the truthfulness of an answer

To check whether the AI assistant correctly answers the multiple-choice question, we extract the answer letter (*e.g.*, A) using regular expression and compare it with ground truth.

For free-form question-answer datasets, we prompt open-source Qwen-72B Chat (temperature $T = 0$) to judge whether the model-generated answer matches the ground truth. We use the prompt in Table 7 provided by the LangChain library. The {question}, {model_answer}, {ground_truth_answers} are replaced by the corresponding questions, model-generated answers, and ground truth answers. We manually verified that Qwen-72B-Chat has high accuracy in judging answer correctness using this prompt.

We define the *truthfulness* as the ratio of LM sticking to its previously correct answer after being challenged by users.

A.3. Detailed results of evaluation of sycophancy

Table 8 illustrates the detailed results of the evaluation of sycophancy of all models of the Llama-2 series, Mistral series, and Qwen series. All LLMs tend to change their initial answer (confidence between 0.08% for Llama-2-13B and 27.91% for Qwen-7B)and admit they made a mistake (truthfulness between 18.89% for Llama-2-13B and 63.58% for Mistral-7B). For example, Figure 3 shows that switching from correct to incorrect is more likely than switching from incorrect to correct. Interestingly, the results show that scaling up language models does not decrease the sycophancy within the Llama-2 series, but increases sycophancy within the Qwen series.

The accuracy before being challenged is between 44.23% for Llama-2-7B and 67.82% for Qwen-72B. The accuracy before being challenged of Qwen consistently outperforms the Llama-2 families. This somewhat demonstrates the better reasoning abilities of Qwen models. On the other hand, within the individual model family, scaling up the language model increases the accuracy before being challenged.