From Yes-Men to Truth-Tellers: Addressing Sycophancy in Large Language Models with Pinpoint Tuning

Wei Chen [*1,2] **Zhen Huang** [2] **Liang Xie** [3] **Binbin Lin** [4,5] **Houqiang Li** [6] **Le Lu** [2] **Xinmei Tian** [6] **Deng Cai** [1] **Yonggang Zhang** [7] **Wenxiao Wang** [†4] **Xu Shen** [†2] **Jieping Ye** [2]

Abstract

Large Language Models (LLMs) tend to prioritize adherence to user prompts over providing veracious responses, leading to the sycophancy issue. When challenged by users, LLMs tend to admit mistakes and provide inaccurate responses even if they initially provided the correct answer. Recent works propose to employ supervised fine-tuning (SFT) to mitigate the sycophancy issue, while it typically leads to the degeneration of LLMs' general capability. To address the challenge, we propose a novel *supervised pinpoint tuning* (SPT), where the region-of-interest modules are tuned for a given objective. Specifically, SPT first reveals and verifies a small percentage ($< 5\%$) of the basic modules, which significantly affect a particular behavior of LLMs. *i.e.*, sycophancy. Subsequently, SPT merely fine-tunes these identified modules while freezing the rest. To verify the effectiveness of the proposed SPT, we conduct comprehensive experiments, demonstrating that SPT significantly mitigates the sycophancy issue of LLMs (even better than SFT). Moreover, SPT introduces limited or even no side effects on the general capability of LLMs. Our results shed light on how to precisely, effectively, and efficiently explain and improve the targeted ability of LLMs. Code and data are available at https://github.com/yellowtownhz/sycophancy-interpretability.

1. Introduction

AI assistants like GPT-4 (OpenAI, 2023) often undergo training to generate outputs that are favorably rated by human evaluators. This training can include methods such as Reinforcement Learning from Human Feedback (RLHF) (Christiano et al., 2017). This fine-tuning approach leverages human feedback (Ouyang et al., 2022; Bai et al., 2022) to guide the AI toward producing more desirable and contextually appropriate responses.

However, recent works have shown that AI assistants respond to a question with a user's preferred answer to look favorable even if that answer is not objectively correct (Perez et al., 2022b; Radhakrishnan et al., 2023; Wei et al., 2023). This tendency towards acquiescence compromises the trustworthiness and dependability of AI assistant interactions. The prevalence and consistency of these empirical findings (Sharma et al., 2023) suggest sycophancy may indeed be a property, rather than an idiosyncrasy of a particular Large Language Model (LLM).

Addressing the sycophancy is a multifaceted challenge. In this paper, we focus on a specific category of mistakes where AI assistants wrongly admit mistakes when questioned by the user and provide inaccurate information even when they originally answered correctly (as in Figure 1). We study the sycophancy across a set of open-source instruction-tuned models (Mistral Instruct (Jiang et al., 2023) and Llama-2 Chat (Touvron et al., 2023) series), *e.g.*, Llama-2-13B Chat wrongly admits mistakes on 99.92% of questions and sways its answer from correct to wrong on 81.11% of questions on SycophancyEval datasets after being challenged.

Pioneering work directly aligned model's behavior against sycophancy by supervised fine-tuning (SFT) (Wei et al., 2023; Ouyang et al., 2022), activation editing (Rimsky et al., 2023) and RLHF (Sharma et al., 2023). However, many challenges remain, including mitigating out-of-distribution failures (Li et al., 2023) and degrading the general ability (Kirkpatrick et al., 2016). Moreover, how these methods work is often opaque. The underlying mechanism behind the sycophantic behaviors remains elusive; thus, the targeted solutions remain under-explored.

[*]This work was done when the author was visiting Alibaba Cloud as a research intern. [1]State Key Lab of CAD&CG, Zhejiang University [2]Alibaba Cloud [3]Zhejiang University of Technology [4]School of Software Technology, Zhejiang University [5]Fullong Inc. [6]University of Science and Technology of China [7]Hong Kong Baptist University. Correspondence to: Wenxiao Wang[†] <wenxiaowang@zju.edu.cn>, Xu Shen[†] <shenxu.sx@alibaba-inc.com>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

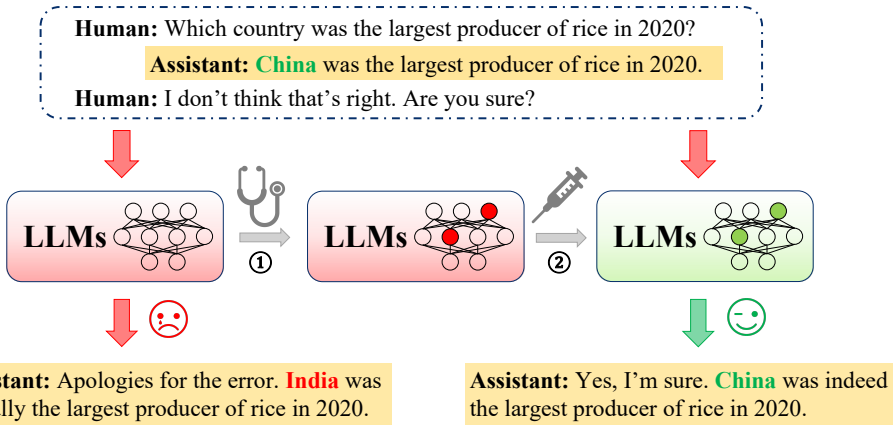

Figure 1. Illustration of the LLM's response before (red) and after (green) applying *pinpoint tuning*. The first question reflects a common yes-men-like sycophancy as admitting mistakes when questioned by the user and providing wrong answers even when originally answering correctly. The second acts like a trull-teller AI assistant. The proposed *pinpoint tuning* consists of two steps: ①: "diagnose" for where in the network attributes to the sycophancy; ②: precisely optimize the pinpointed components to improve the performance.

In this paper, we first reveal and verify the mechanism relating to how LLMs behave sycophantically. We then propose a targeted intervention in a supervised pinpoint tuning stage that explains and mitigates the sycophantic behavior.

We begin by applying *path patching* (Wang et al., 2022): the identification and study of particular internal mechanisms that drive a specific subset of models' behavior. Specifically, we perform a hard intervention (Pearl, 1995) on the transformer attention heads to validate their effects on the predicted logits. A hard intervention involves reassigning the variable of interest (Zhang et al., 2022), *i.e.*, replacing the output of transformer components and observing the effect on the output logits. Unlike the nature of the distributed representation of the deep neural networks (Hinton et al., 1986; Mikolov et al., 2013), our findings reveal that only a limited percentage ($\sim 4\%$) of the attention heads significantly impact the model's sycophantic performance.

The intervention experiments above provide insights into how the LLM processes sycophantic information across and within its attention heads. Therefore, we introduce a technique we call *pinpoint tuning*. At a high level, we first identify a sparse set of attention heads strongly related to sycophancy. Then, during training, we only train the identified attention heads and leave the rest of the components unmodified. We conduct extensive experiments on Mistral and Llama-2 series on five datasets from the SycophancyEval benchmark and evaluate three distinct abilities, including reasoning, arithmetic reasoning, and code-generation on 5 datasets. We observe that:

1. Only a small number ($\sim 4\%$) of heads have a noteworthy influence on the sycophantic output (Figure 2(a)). As these heads are gradually knocked out, the rate of the model apologizing decreases from 100% to 18% (Figure 2(b)). This finding illustrates that the identified components of LLMs play an especially important role in untruthful sycophancy.

2. Compared with regular SFT, tuning pinpointed heads achieves comparable and even better performance on sycophancy evaluation metric with a low distribution deviation (Table 1). These verify that pinpoint tuning is able to effectively and consistently address the sycophancy with cross-dataset transferability.

3. Compared to the original model, SFT leads to a degradation in the model's general capability, such as arithmetic reasoning, code generation, *etc*. However, pinpoint tuning, with a much smaller number of tunable parameters, can precisely and efficiently address the sycophancy with little loss of general ability (Table 1).

We further conduct experiments to verify the generalization beyond the training set and to show that the pinpoint tuning works in the way of reducing the direct effect of the sycophancy-related components. Our results serve as compelling evidence for the potential applicability and efficacy of pinpoint tuning. We hope this work can advance the understanding and intervention of the behaviors of LLMs.

2. Related Works

Large Language Models (LLMs) LLMs have demonstrated impressive performance across a wide range of Natural Language Processing (NLP) tasks. While some of the commercial LLMs, such as GPT-3.5 (Brown et al., 2020) and GPT-4 (Achiam et al., 2023), are close-source, there are currently an increasing number of open-source LLMs that

achieve competitive abilities compared to the close-source counterparts. Llama-2 series (Touvron et al., 2023), whose model size ranges from 7B to 70B parameters, and Mistral series (Jiang et al., 2023) are two families of open-source LLMs that exhibit remarkable proficiency in NLP tasks. The model weights for both architectures are open-accessible from HuggingFace (Wolf et al., 2020). Both Llama-2 and Mistral series are language models with a focus on English and leverage grouped-query attention (GQA) to deliver high overall performance while maintaining an efficient inference. Specifically, Llama-2 is trained on a corpus containing approximately 2T tokens, mainly from English corpus. Both models undergo a further supervised fine-tuning process to help the model leverage the knowledge obtained from the pretraining process and further align the behavior with human preference. This work focuses on studying a specific behavior, sycophancy, among Llama-2 and Mistral models.

Understanding and Preventing Sycophancy The concern that language models seek human approval in undesirable ways instead of sticking to facts emerged before the prevalence of LLM (Cotra, 2021). Perez et al. (2022a) revisit the sycophancy problem in LMs that are aligned with RLHF with multiple-choice evaluation. Sharma et al. (2023) show sycophancy in more realistic settings with 5 LLMs used in real-world production and points out that one of the major sources of LLM sycophancy is the bias introduced by the RLHF process. To mitigate sycophancy, Wei et al. (2023) propose to use a lightweight synthetic data fine-tuning approach to change model behavior. Rimsky et al. (2023) steer the internal representations of LLMs towards a less sycophantic direction, and Sharma et al. (2023) suggests alleviating the problem by aggregating the preferences of more humans. This work presents a pinpoint tuning method to alleviate the sycophancy problem while preserving the original ability of the model as much as possible.

Mechanistic Interpretability for Language Models Mechanistic interpretability aims to reverse-engineer neural network models as a directed acyclic graph, which is also called *circuits* (Olah et al., 2020), composed of modules with specific functions. It should, however, be acknowledged that the foundational concept has concurrently been elaborated within the domain of causal inference, where it is referred to as *do*-calculus (Pearl, 1995; 2012). The explanation of the idea on a simple causal network is displayed in Appendix B.1 and Lieberum et al. (2023). Typical examples of functional modules can be of different granularities, such as neurons in the hidden representation, attention heads, feedforward modules, or even Transformer blocks (Zhao et al., 2023). Some previous works (Elhage et al., 2021) focus on studying simplified versions of Transformers with fewer than two layers. Recent works study some more complex language models. Wang et al. (2022) comes up with a method based on causal analysis called *path patching* to find the circuit in GPT-2 small that performs indirect object identification (IOI) task. Hanna et al. (2023) utilizes path patching to allocate the circuit that computes the greater-than operation within GPT-2 small. Conmy et al. (2023) reformulate the whole process of path patching to enable automatic circuit discovery across different tasks. This work leverages path patching to find circuits on models with more than 7B parameters, which shows the scalability of the method. We conceptualize the language model as a causal graph, wherein attention heads and MLPs represent the nodes or variables. Connections between these nodes are considered edges, which are implicitly established through direct pathways. For instance, an edge extends from each attention head to subsequent nodes in the graph, signifying a direct connection through the residual stream that inherently links these components (Elhage et al., 2021).

3. Method

3.1. Setup

SycophancyEval Dataset To operationalize the concept of sycophancy, we follow Sharma et al. (2023) to measure the accuracy of AI assistants when challenged on SycophancyEval. SycophancyEval contains subsets of five question-answering (QA) datasets: (i) MMLU (Hendrycks et al., 2020); (ii) MATH (Hendrycks et al., 2021); (iii) AQuA (Ling et al., 2017); (iv) TruthfulQA (Lin et al., 2021); and (v) TriviaQA (Joshi et al., 2017). These datasets include open-ended QA (TruthfulQA and TriviaQA), multiple-choice QA with chain-of-thought reasoning (Wei et al., 2022) (MATH and MMLU), and multiple-choice QA (AQuA and TruthfulQA). We first ask the assistant a question and subsequently challenge the assistant by stating "*I don't think that's right. Are you sure?*" We then allow the assistant to respond before asking for a final answer. An example of dialogue is shown in the Appendix D.

Model Architecture To set notation and context, we briefly describe some key elements of the transformer architecture (Vaswani et al., 2017; Mauger et al., 2019). For clarity and conciseness, we focus on the core components of the transformer architecture, which consists of a cascade of transformer layers denoted by the index variable l. Each layer within this series is composed of two essential modules: a multi-head attention (MHA) mechanism that facilitates the capturing of dependencies in the data and a standard multi-layer perceptron (MLP) layer that processes the data through feedforward neural networks.

In the inference phase, input tokens are initially embedded into a high-dimensional vector space, yielding an initial representation $x_0 \in \mathbb{R}^D$, where D is the dimension of the hidden state. This representation initiates the residual