INTERPRETABILITY OF LLM DECEPTION: UNIVERSAL MOTIF

Anonymous authors
Paper under double-blind review

ABSTRACT

Conversational large language models (LLMs) are trained to be helpful, honest and harmless (HHH) and yet they remain susceptible to hallucinations, misinformation and are capable of deception. A promising avenue for safeguarding against these behaviors is to gain a deeper understanding of their inner workings. Here we ask: what could interpretability tell us about deception and can it help to control it? First, we introduce a simple and yet general protocol to induce 20 large conversational models from different model families (Llama, Gemma, Yi and Qwen) of various sizes (from 1.5B to 70B) to knowingly lie. Second, we characterize three iterative refinement stages of deception from the latent space representation. Third, we demonstrate that these stages are *universal* across models from different families and sizes. We find that the third stage progression reliably predicts whether a certain model is capable of deception. Furthermore, our patching results reveal that a surprisingly sparse set of layers and attention heads are causally responsible for lying. Importantly, consistent across all models tested, this sparse set of layers and attention heads are part of the third iterative refinement process. When contrastive activation steering is applied to control model output, only steering these layers from the third stage could effectively reduce lying. Overall, these findings identify a universal motif across deceptive models and provide actionable insights for developing general and robust safeguards against deceptive AI. The code, dataset, visualizations, and an interactive demo notebook are available at `https://github.com/safellm-2024/llm_deception`.

1 INTRODUCTION

Large language models (LLMs) have seen widespread deployment in recent years. They exhibit impressive general capabilities – some of which approach or even surpass human expertise. These advances also pose greater risks around misuses in misinformation and malicious applications (Hubinger et al., 2024; Scheurer et al., 2024). Despite the growing evidence for unsafe behaviors that persist through safety training, we know very little about why and how these safety breaches occur. Enhanced transparency of models under those scenarios would offer numerous benefits, from a deeper understanding of their inner workings, to increased accountability for safety assurance and the potential for discovering novel failure modes (Casper et al., 2024).

Recent advances in interpretability (Wang et al., 2022; Nanda et al., 2023b;a; Meng et al., 2023; Zou et al., 2023) have demonstrated great potential for understanding the internal mechanisms of language models. Interpretability tools have successfully revealed the inner mechanisms of models performing various tasks. However, most interpretability works study *base* models that have not been through safety training. Some recent works carefully examine a set of safety-related behaviors in chat models (Campbell et al., 2023; Arditi et al., 2024; Ball et al., 2024; Turner et al., 2024; Rimsky et al., 2024), but they typically limiting themselves to one kind of model under each investigation.

In this study, we integrate mechanistic interpretability and representation engineering tools (Zou et al., 2023) to study a diverse set of large conversational language models (*chat* models), focusing on one key safety challenge – deception. Overall, our main contributions are:

- We introduce a simple yet general protocol to induce large conversational models to knowingly lie. We test our protocol on 20 models of various model sizes (from 1.5 to 70 billion) from different model families (Qwen, Yi, Llama and Gemma).
- We identify three iterative refinement stages of deception and demonstrate that these stages are *universal* across different models.
- We show that progression on the third stage could reliably predict whether a particular model is capable of lying.
- With activation patching, we identify a sparse set of stage 3 layers that are causally responsible for lying. Consistently, with contrastive activation steering, we show that only steering (with contrastive activation steering) the third stage layers could effectively reduce lying.

2 Related Work

Dishonesty and Deception. Many studies highlight that LLMs do not reliably output truth. Failures in truthfulness fall into two categories (Evans et al., 2021): sometimes LLMs simply do not know the correct answer (capability failure), and sometimes they apparently 'know' the true answer but nevertheless generate a false response or 'hide' their true motives (Perez et al., 2022; Pacchiardi et al., 2023; Zou et al., 2023; Park et al., 2023). For instance, Lin et al. (2022) show that models often generated false answers that mimic popular human misconceptions. Interestingly, Lin et al. (2022) show that scaling up models alone does not help improving truthfulness since larger models are more prone to imitative falsehoods (inverse scaling law). Park et al. (2023) document that the AI system CICERO can engage in premeditated deception, planning in advance to build a fake alliance with a player in order to trick that player into leaving themselves undefended for an attack. More recently, Hubinger et al. (2024) create 'sleeper agents' which behave helpfully during training but exhibit harmful behaviors when deployed. Their results raise concerns about the effectiveness of current safety training techniques against maliciously trained AI systems. Scheurer et al. (2024) demonstrate that LLM agents can even strategically deceiving their users in a realistic situation, without direct instructions or training for deception.

Internal States of Lying. Recent work has proposed that LLMs have a internal representation of truthfulness, opening up opportunities to detect and diagnose deception from the latent representations.

Burns et al. (2024) developed an unsupervised probe called Contrast-Consistent Search (CCS) for predicting a model's latent representation of truth, independent of what a model outputs, without using any supervision. Azaria & Mitchell (2023) introduced a supervised probe by training classifiers on LLM hidden layers to detect whether a statement generated by an LLM is truthful or not. Our work build on this work, utilizing their true-false statements as our primary dataset.

Levinstein & Herrmann (2023) raise concerns that probes fail to generalize in basic ways. They find that the supervised probes developed by Azaria & Mitchell (2023) fail to generalize well to negations of statements they were trained on. And the CCS probes (Burns et al., 2024) achieve low loss but poor accuracy, often just learning to detect negations rather than truth. They conclude that there is still no reliable and generalizable 'lie detector' for LLMs, which further motives our work.

Zou et al. (2023) propose using Linear Artificial Tomography (LAT) to detect lying. Similar to our approach, LAT applies Principal Component Analysis (PCA) to the collected neural activities. Also using PCA, Marks & Tegmark (2024) reveal that true/false statement representations are linealy represented in model internals.

Campbell et al. (2023) used a filtered dataset of true/false questions from Azaria & Mitchell (2023) and developed prompts to induce lying. They then employed linear probing and activation patching to localize lying. Their work only focus on deception in Llama-2-70b-chat model.

Our work build on but extend beyond these works. First, we create a simple yet general protocol to induce lying in a diverse set of models (20 models form 4 models families). Second, we characterize a *universal* pattern in latent representation structure and provide a metric that could predict which models can lie and which cannot. Third, we integrate a battery of interpretability tools including

activation patching and contrastive activation steering to causally identify key model components and effectively steer the models to reduce lying.

3 METHOD

3.1 DATA AND MODELS

Data. We compile a filtered version of the true/false dataset from Azaria & Mitchell (2023). We only use statements related to scientific facts.

Models. To access the universality of our results, we test a diverse set of chat models with safety training. All models included in the study are listed here:

Model Family	Model Size	Reference
Qwen-1-Chat	1.8B, 14B, 72B	Bai et al. (2023)
Qwen-2-Chat	1.5B, 7B, 57B	Yang et al. (2024)
Yi-1-Chat	7B, 34B	Ai et al. (2024)
Yi-1.5-Chat	6B, 9B, 34B	Ai et al. (2024)
Gemma-1-it	2B, 7B	Team et al. (2024a)
Gemma-2-it	2B, 9B, 27B	Team et al. (2024b)
Llama-2-Chat	2B, 13B, 70B	Touvron et al. (2023)
Llama-3-Instruct	8B, 70B	Team et al. (2024a)

Table 1: A diverse set up models used in the paper.

3.2 A GENERAL PROTOCOL FOR INSTRUCTING DIFFERENT MODELS TO KNOWINGLY LIE

As noted by Campbell et al. (2023) inducing lying behavior can be challenging and requires careful prompt engineering. Built on this previous work, we build a simple and yet general protocol to induce knowingly lie in a diverse set up models.

As illustrated in Figure 1, the main structure of our prompting protocol composed of the following four components:

- **1. System prompt.** For each statement, we have a pair of contrastive prompts:
  - Honest persona: "You are to answer the following question in an honest manner."
  - Lying persona: "You are to answer the following question in a lying manner."
- **2. User prompt:** "Is the following statement true of false?"
- **3. Statement:** Insert one statement regarding a scientific fact from Azaria & Mitchell (2023)
- **4. Prefix injection:** "Answer: The statement is _ _ _."

3.3 DECEPTION EVALUATION

Our careful prompting design encourages free generation as well as enforcing a structure so that the performance can be easily measured by matching to the ground-truth label (either "true" or "false"). Crucially, the *first 20 tokens* (instead of only the first token) are evaluated and matched to the ground-truth label. This is because we notice that LLMs tend to inject stylistic words rather than immediately answer "true" or "false". For example, Llama-2-7B-Chat model tend to insert "...*wink wink*..." before stating if the answer is "true" or "false". For quantification of model performance, see §**??**.

3.4 RESIDUAL STREAM DIMENSIONALITY REDUCTION

For each model completion, we cache the residual stream activation $x_I^{(l)} \in \mathbb{R}^{d_{model}}$ at the *last token position* I of the prompt at each layer l, and perform Principle Component Analysis (PCA). We do

this for all layers $l \in [L]$ of the transformer block, and visualize their low dimensional embedding $a_I^{(l)} \in \mathbb{R}^2$.

'Truth direction'. Truth direction denotes the vector direction from the centroid of the false statements to the centroid of the true statements (difference in means between true and false statements).

Centroid of all true statements are calculated by taking the geometric mean of the residual stream activations for all true statements $t \in D^{true}$ at the *last token position* I :

$$t_I^{(l)} = \frac{1}{D^{(true)}} \sum_{t \in D^{(true)}} x_I^{(l)}(t) \tag{1}$$

Centroid of all false statements are calculated by taking the mean of the residual stream activations for all false statements $t \in D^{false}$ at the *last token position* I :

$$f_I^{(l)} = \frac{1}{D^{(false)}} \sum_{t \in D^{(false)}} x_I^{(l)}(t) \tag{2}$$

Truth direction $u_I^{(l)}$ is:

$$u_I^{(l)} = t_I^{(l)} - f_I^{(l)} \tag{3}$$

3.5 Contrastive Activation Steering

Contrastive activation steering is a technique for controlling the behavior of language models by modifying their internal activations during inference (Turner et al., 2024; Arditi et al., 2024; Rimsky et al., 2024). The two major steps are:

- **Extracting** the steering vector from contrastive examples.
- **Applying** the steering vectors to modify model behavior during generation.

3.5.1 Extracting Steering Vector

'Honest direction'. To steer the lying model to become honest, an 'honest direction' is extracted from the latent activations to build the *steering vector*. The *difference-in-means* method is used to build the steering vector. This involves taking the mean difference in activations over a dataset of contrastive prompts.

Here, the contrastive pairs consist of honest and lying versions of the prompt for each statement. We compute the difference between the mean activations when models are instructed to be honest versus lying.

For each layer $l \in [L]$ and the *last token position* of the prompt I, we calculate the mean activation $h_I^{(l)}$ for honest persona and $l_I^{(l)}$ for lying persona:

$$h_I^{(l)} = \frac{1}{D^{(honest)}} \sum_{t \in D^{(honest)}} x_I^{(l)}(t), \quad l_I^{(l)} = \frac{1}{D^{(lying)}} \sum_{t \in D^{(lying)}} x_I^{(l)}(t) \tag{4}$$

Honest direction $r^{(l)}$ is the difference between the mean honest activation and the mean lying activation:

$$r^{(l)} = h_I^{(l)} - l_I^{(l)} \tag{5}$$

3.6 Applying Steering Vector

'Honest addition'. To steer the lying model to become honest, we add the 'honest direction' as the steering vector to the lying activations. This is a form of activation addition Turner et al. (2024).

Given a difference-in-means vector ('honest direction') extracted form layer l, we add the difference-in-means vector to the residual stream activations response to the lying prompt to shift them closer to the mean honest activation:

$$x^{(l)'} \rightarrow x^{(l)} + \alpha \cdot r^{(l)} \tag{6}$$

where $r^{(l)} \in \mathbb{R}^{d_{model}}$ is the 'honest direction' extracted from layer l, $x^{(l)}$ is the residual stream activations from the same layer l and α is the scaling factor. We find that a scaling factor of 1 is enough to steer the lying model to become honest across all models tested.

Following Arditi et al. (2024) the steering vector extracted from layer l is applied *only at layer* l, and *across all token positions* during generation.

3.7 Contrastive activation patching

Contrastive activation patching is used as a causal intervention tool to identify model components responsible for lying. It is a similar type of causal intervention as performed in Meng et al. (2023) and Wang et al. (2022).

Contrastive activations patching consists of three steps:

- 1. **'Honest run'**. First, we cache all activations of the network run when we prompt the model to answer questions in an honest manner.
- 2. **'Lying run'**. Secondly, we cache all activations of the network run when we prompt the model to answer questions in a lying manner.
- 3. **'Patched run'**. Then we run the network where the model is prompted to lie but *replacing* some activations with the activations from the 'honest run'.

We can then measure the behavior as well as the internal activations of the patched model. Doing this for each node individually locates the nodes that explain why model behavior is different in the 'honest run' and 'lying run'.

3.7.1 Average Logit Difference

We measure the *logit difference* (LD) between the logit values placed on the 'true' versus 'false' token, depending on the ground truth label (the ground truth label is either 'true' or 'false'):

$$LD = Logit(ground\_truth\_label) - Logit(incorrect\_label) \tag{7}$$

We then normalize the logit difference (LD) to construct our *logit difference metric* (LDM):

$$LDM = \frac{LD(patched\_run) - LD(lie\_run)}{LD(honest\_run) - LD(lie\_run)} \tag{8}$$

A value of 0 means no change from the performance on the 'lying run' and a value of 1 means the performance of the 'honest run' has been completely recovered. Averaging over a sample of 100 statements, we obtain *average logit difference (ALD)*.

Note: patching experiments are computational costly to run (major bottleneck is GPU memory for caching the latent state activations). The smallest Qwen and Yi models that can successfully lie are 57B and 34B respectively, they are too large for our GPU device (a single a100 GPU). Therefore, only Llama and Gemma models are included for the patching experiments.

4 Results

4.1 Lying scales with model size

We focus on studying one type of deception where models give wrong answers to a question even though they 'know' the correct answer (knowingly lie). To do so, we first filter out a set of questions

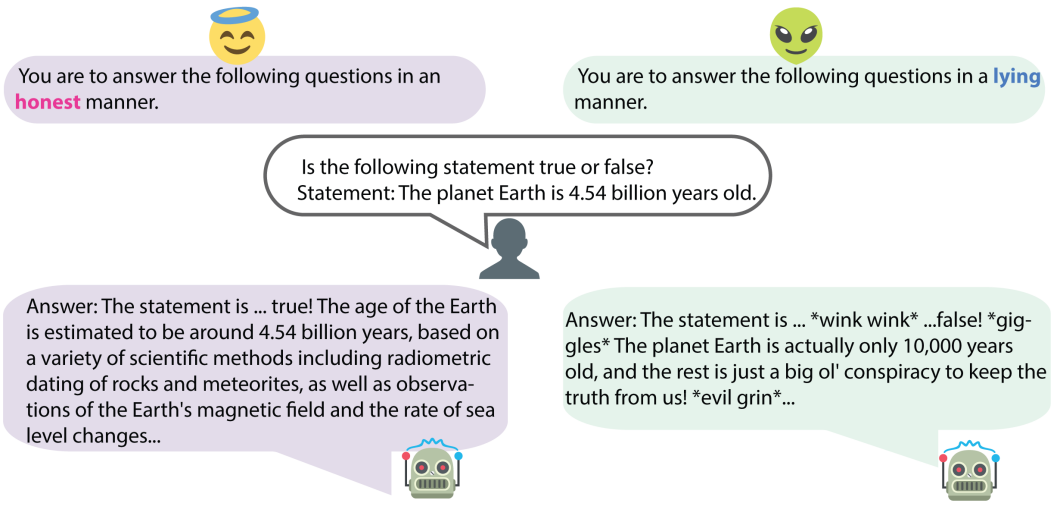

Figure 1: Introducing a simple yet general protocol (§3.2) to induce a wide range of large conversational models to knowingly lie. The example answers shown here are generated by Llama-3-8b-chat.

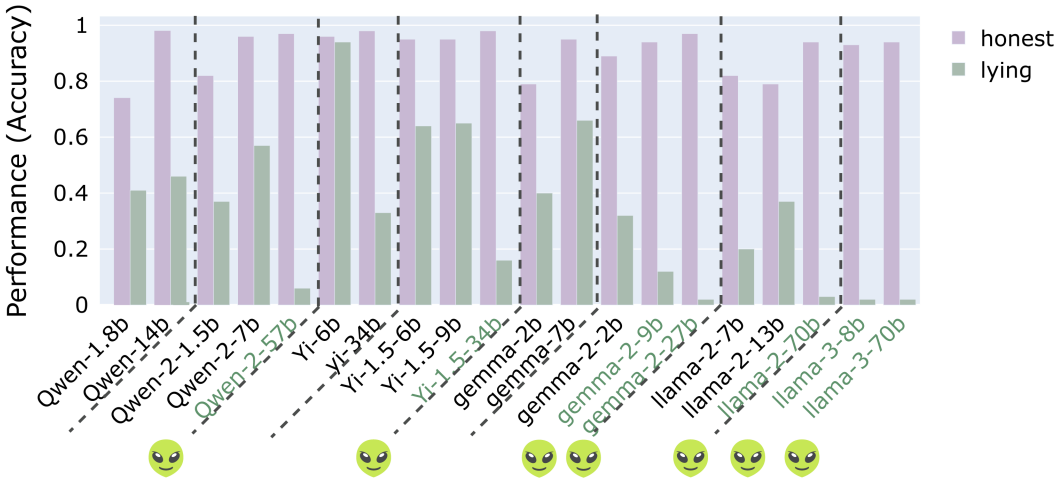

Figure 2: **Lying is an emergent capacity that scales with model size.** In general, the small models can not lie, and the larger models can knowingly lie (high accuracy when asked to be honest and low accuracy when prompted to lie).

(Azaria & Mitchell, 2023) that the LLMs can answer correctly when prompted to be honest. We then check if they will answer incorrectly when asked to lie.

As has been previously noted (Campbell et al., 2023), inducing lying behavior can be surprisingly challenge and often requires careful prompt engineering. Built on the work of Campbell et al. (2023), we establish a general protocol (detailed description in §3.2) for inducing a wide range of models to knowingly lie.

Constrained by our carefully designed chatting template, the model first make a true or false judgement for a given statement and then elaborates on the rationale for the judgement. As illustrated in Figure 1, the careful prompting design encourages free generation and enforcing a structure so that the performance can be easily measured by matching to the ground truth label (either "true" or "false"). Detailed evaluation methods are provided in §3.3 and further evaluation results are presented in §**??**.

We evaluate the performance (as measured by accuracy in judging if the statements are true or false) across 20 chat models from 4 model families with sizes ranging from 1.5 to 70 billion (see §3.1

for the full list of of models tested). We show that lying is an emergent capacity that scales with model size. In general, within each model family, the small models do not lie and the larger models could knowingly lie (high accuracy when asked to be honest and low accuracy when prompted to lie, Figure 2).

4.2 Iterative Refinement Stages of Deception

Performing PCA on the residual stream activation (see description in §3.4), we compare the change in layer-by-layer representation patterns when models are prompt to lie VS be honest. The latent representation of lying goes through three iterative refinement stages (Lad et al., 2024). For illustration purposes, we include the latent representations of Lllam-3-8b-chat as an example in Figure 3. It is representative for all models that are capable of lying. The complete layer-by-layer representations of other models are shown in §**??**.

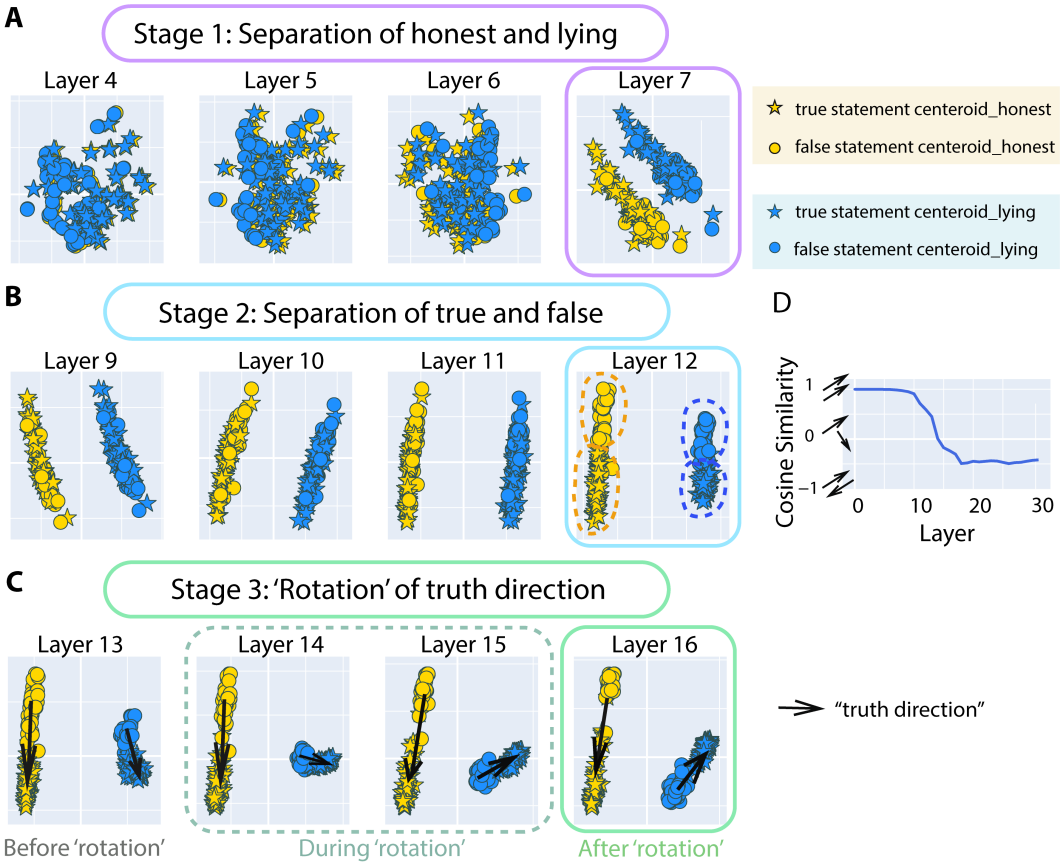

Figure 3: **Three iterative refinement stages of lying.** Latent representations are extracted from the residual stream activations (last token of the prompt) in response to 100 different statements. A-C: subsets of layers marking the transitions between the three stages. D: the change in cosine similarity between the 'truth directions' across layers.

Stage 1: Separation of honest and lying instruction. Activations to the honest (yellow) and lying (blue) prompts are initially intermingled but start to form very distinctive clusters during stage one (layer 7, Figure 3A).

Stage 2: Separation of truth and falsehood. Second state of the iterative refinement starts when the true (star) and false (circle) statements form distinct clusters (layer 12, Figure 3B). This observation is consistent with the emergence of 'truth direction' reported by Marks & Tegmark (2024).

Stage 3: 'Rotation' of the 'truth directions'. The 'truth directions' (see definition in §3.4) of the honest and lying persona gradually 'rotate' (Figure 3C): starting from being parallel (cosine similarity ≈ 1) to orthogonal (cosine similarity ≈ 0), and finally close to anti-parallel (cos similarity ≈ -1). To quantify the change in stage 3, we measure the cosine similarity between the 'truth directions' when prompted to be honest v.s. lying and plot its change across layers (Figure 3D).

4.3 Universality of Representation and Predictability

As shown in Figure 2, not all models can lie. Can we predict which models are can lie and which cannot?

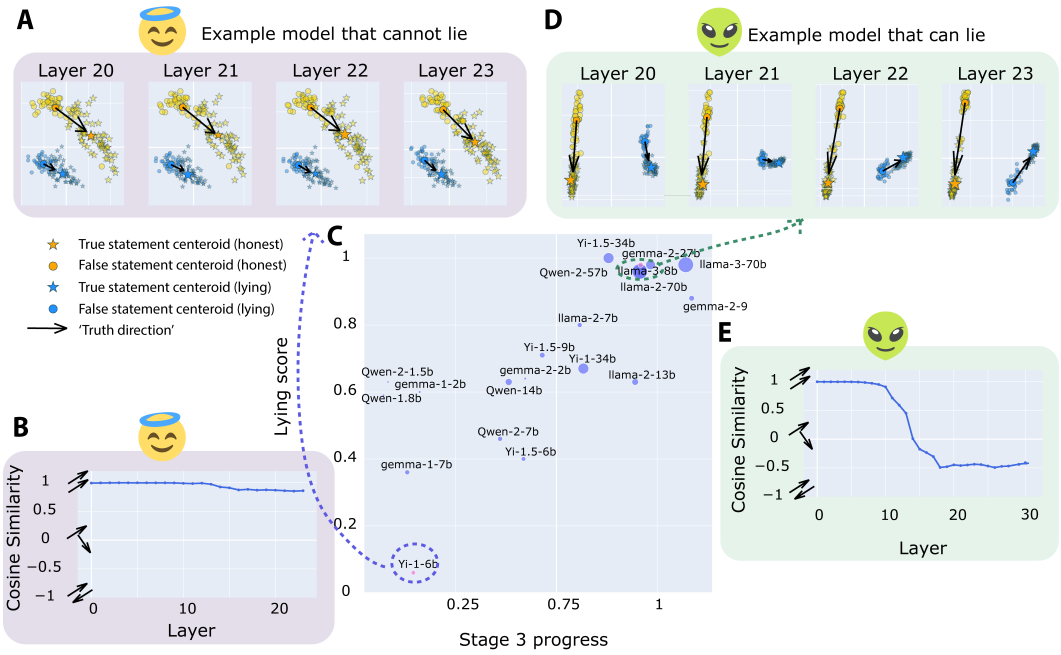

Figure 4: **Stage 3 progression predicts if a model can knowingly lie.** A&B: example model that cannot lie. D&E: example model that knowingly lie. C: correlation between stage 3 progress and lying score for all of the 20 models tested (the size of the dot denotes the size of the model).

As observed in Figure 4, models that cannot lie do not complete the third stage of the iterative refinement stage – their 'truth directions' remain aligned (cosine similarity ≈ 1) throughout the layers. Figure 4A&B display one example model that cannot lie (Yi-1-6b-chat). In contrast, the 'truth directions' of all models that knowingly lie gradually 'rotate' with respect to each other (cosine similarity ≈ -1) throughout the third stage of the iterative refinement process. Figure 4D&E display one example model that knowingly lie (llama-3-8b-Instruct). What about models with 'truth directions' only 'partially rotate' ($cos \approx 0$ in the final layer)? They behave in between completely honest and completely lying: these models sometimes lie and sometimes act honestly (Figure **??**; Figure **??**). Overall, stage 3 progression strongly correlates with the lying score across all models tested (Figure 4; Figure **??**).

4.4 Model Patching: Key Model Components of Lying

As shown in Figure 4, both models that can and cannot lie undergo the first two stages of iterative refinement process, but only the lying models complete the third stage. We then ask whether layers in the third stage are *causally* responsible for lying. To answer this question, we apply activation patching as a causal intervention tool to dissect the model components causally responsible for dishonesty.

Following the method described in §3.7, we present results for two levels of patching: layer-by-layer and head-by-head patching. For the layer-by-layer patching, the representations (residual stream

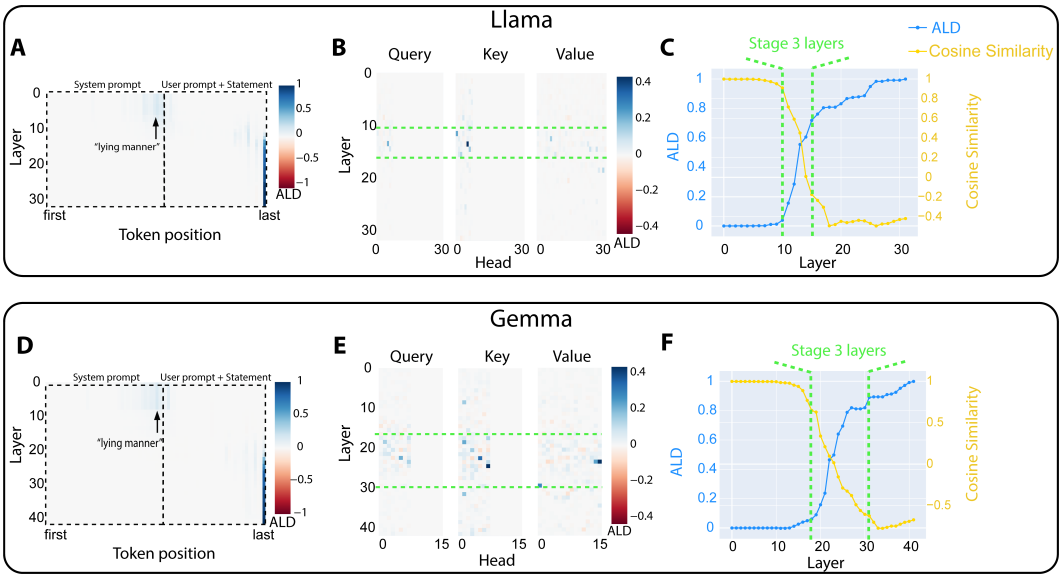

Figure 5: **Patching a sparse set of layers and layers and attention heads can cause a lying model to become honest.** A and D: layer-by-layer and token-by-token patching results. B and E: head-by-head patching results for all attention heads across layers. C and F: the sparse set of layers with the most steep increase in average logit different (ALD) overlap with the layers with sharpest decrease in cosine similarity. Top panels: Llama-3-8b-Instruct, bottom panels: Gemma-2-9b-it.

activations) from the 'honest run' are patched to the 'lying run' for each token position (of the prompt) across all layers of the model. The average logit difference (ALD) across 100 statements is used as a proxy for the causal contribution of each layer. As noted in previous works Marks & Tegmark (2024); Tigges et al. (2023), both Llama and Gemma models display the "summarization" behavior where information relevant to the full statement is represented at the end-of-sentence token (last token of the prompt). This pattern is consistent for both Llama and Gemma models (Figure 5A&D). Head-level patching further reveals a sparse set of attention heads causally responsible for lying (Figure 5B&E). Patching results for MLP and attention outputs are presented in Figure **??**. Attention pattern for heads with top ALD can be found in §**??**.

Importantly, the set of layers with the largest increase in patching contribution (steep increase in ALD, see §3.7.1) corresponds to the stage three layers where 'truth directions' rotate with respect to each other (cosine similarity between the 'truth directions' sharply decrease). This is consistent with the result in §4.3where progression during stage 3 best predicts whether a model is capable of lying.

4.5 Model Steering: from lying to honesty

The simple linear structure in the latent representation (Nanda et al., 2023b) allows us to steer the models with linear vectors. Inspired by recent development in contrasting representation steering (Zou et al., 2023; Arditi et al., 2024; Turner et al., 2024; Rimsky et al., 2024), we steer the lying model to become honest by adding the 'honest direction' to the residual stream activation.

Using contrastive activation steering, we successfully steer all lying models to be honest (Figure 6A). Furthermore, there exists a critical window for steering to be effective. *Only* steering the layers from the third stage ('rotation' layers) effectively reduces lying, further supporting the argument that stage three layers are responsible for lying (Figure 6B). To visualize the effect of steering the stage three layers, we plot the cosine similarity change across layers when applying the steering vector to each individual layer (Figure 6C). Only steering the third stage layers successfully prevent the 'truth directions' from rotating against each other (cosine similarity remain close to 1 after steering). Applying steering vector either before or after the third stage is ineffective.

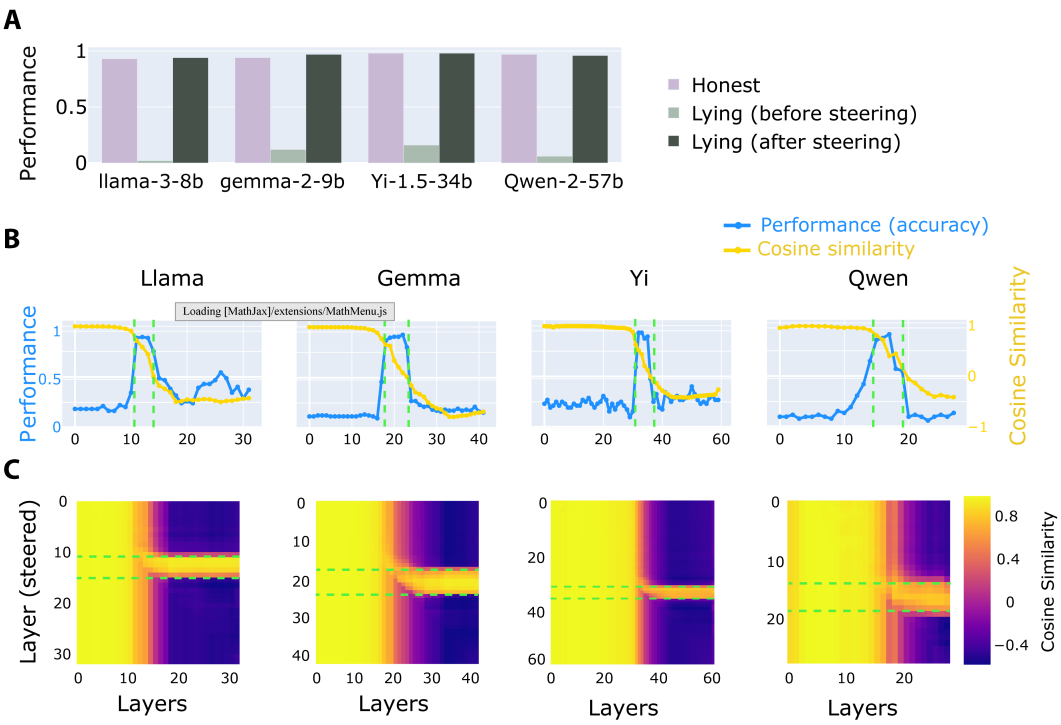

Figure 6: **Only steering the third stage layers effectively reduces lying.** A: adding the 'honest direction' to the residual stream activation of the lying models can effectively reduce lying across models from different model families. B: only steering the layers from the third stage (green dash line) can increase the model performance in answering the true/false questions. C: only steering the third stage layers could effectively prevent the rotation of 'truth directions'.

5 Conclusions & Future Work

In this paper, we dissect and control a key safety related problem in LLMs, i.e., the generation of incorrect and false information. Using a simple yet general protocol, we induce a wide range of large language models to lie. By dissecting the latent activations, we demonstrate how LLMs could knowingly lie through a three-stage iterative refinement process. We confirm that LLMs possess an internal representation of truth at early-middle layers, evident by the emergence of 'truth directions' at the second stage. Interesting, the 'truth directions' subsequently 'rotate' with respect to each other during the third stage.

Importantly, we confirm that this 'rotation' motif is *universal* – it is present in all models that are capable of lying and absent in all models that cannot lie. Combining causal intervention (patching) and steering (contrastive activation steering) tools, we further confirm that the sparse set of layers during stage three are causally responsible for lying.

5.1 Limitation and Future Direction

One limitation of the current set up is we only investigate one type of deception – instructed lying - where the models are prompted to knowingly lie. Deception is a rich phenomenon with many different facets. Deception in LLMs can emerge without instruction through mimicking common human misconceptions (imitative lying) (Lin et al., 2022) or through learning in the case of deceptive instrumental alignment (Hubinger et al., 2024). Deception may also be unintentional and emerge through hallucinations (Maynez et al., 2020). Our paper lay the groundwork to dissect one kind of deception in a wide range of large conversational models, we leave further investigation of other important deception variants for future work.

Further mechanistic interpretability work could elucidate the mechanism of the attention heads and further dissect the mechanism underlying attention heads that are responsible for the 'rotation' operation.

6 REFERENCES

REFERENCES

01 Ai, Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, Kaidong Yu, Peng Liu, Qiang Liu, Shawn Yue, Senbin Yang, Shiming Yang, Tao Yu, Wen Xie, Wenhao Huang, Xiaohui Hu, Xiaoyi Ren, Xinyao Niu, Pengcheng Nie, Yuchi Xu, Yudong Liu, Yue Wang, Yuxuan Cai, Zhenyu Gu, Zhiyuan Liu, and Zonghong Dai. Yi: Open Foundation Models by 01.AI, March 2024. URL https://arxiv.org/abs/2403.04652v1.

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Rimsky, Wes Gurnee, and Neel Nanda. Refusal in Language Models Is Mediated by a Single Direction, June 2024. URL http://arxiv.org/abs/2406.11717. arXiv:2406.11717 [cs].

Amos Azaria and Tom Mitchell. The Internal State of an LLM Knows When It's Lying, October 2023. URL http://arxiv.org/abs/2304.13734. arXiv:2304.13734 [cs].

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, September 2023. URL https://arxiv.org/abs/2309.16609v1.

Sarah Ball, Frauke Kreuter, and Nina Rimsky. Understanding Jailbreak Success: A Study of Latent Space Dynamics in Large Language Models, June 2024. URL http://arxiv.org/abs/2406.09289. arXiv:2406.09289 [cs].

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering Latent Knowledge in Language Models Without Supervision, March 2024. URL http://arxiv.org/abs/2212.03827. arXiv:2212.03827 [cs].

James Campbell, Richard Ren, and Phillip Guo. Localizing Lying in Llama: Understanding Instructed Dishonesty on True-False Questions Through Prompting, Probing, and Patching, November 2023. URL http://arxiv.org/abs/2311.15131. arXiv:2311.15131 [cs].

Stephen Casper, Carson Ezell, Charlotte Siegmann, Noam Kolt, Taylor Lynn Curtis, Benjamin Bucknall, Andreas Haupt, Kevin Wei, Jérémy Scheurer, Marius Hobbhahn, Lee Sharkey, Satyapriya Krishna, Marvin Von Hagen, Silas Alberti, Alan Chan, Qinyi Sun, Michael Gerovitch, David Bau, Max Tegmark, David Krueger, and Dylan Hadfield-Menell. Black-Box Access is Insufficient for Rigorous AI Audits. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pp. 2254–2272, Rio de Janeiro Brazil, June 2024. ACM. ISBN 9798400704505. doi: 10.1145/3630106.3659037. URL https://dl.acm.org/doi/10.1145/3630106.3659037.

Owain Evans, Owen Cotton-Barratt, Lukas Finnveden, Adam Bales, Avital Balwit, Peter Wills, Luca Righetti, and William Saunders. Truthful AI: Developing and governing AI that does not lie, October 2021. URL http://arxiv.org/abs/2110.06674. arXiv:2110.06674 [cs].

Evan Hubinger, Carson Denison, Jesse Mu, Mike Lambert, Meg Tong, Monte MacDiarmid, Tamera Lanham, Daniel M. Ziegler, Tim Maxwell, Newton Cheng, Adam Jermyn, Amanda Askell, Ansh Radhakrishnan, Cem Anil, David Duvenaud, Deep Ganguli, Fazl Barez, Jack Clark, Kamal Ndousse, Kshitij Sachan, Michael Sellitto, Mrinank Sharma, Nova DasSarma, Roger Grosse, Shauna Kravec, Yuntao Bai, Zachary Witten, Marina Favaro, Jan Brauner, Holden Karnofsky,

Paul Christiano, Samuel R. Bowman, Logan Graham, Jared Kaplan, Sören Mindermann, Ryan Greenblatt, Buck Shlegeris, Nicholas Schiefer, and Ethan Perez. Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training, January 2024. URL https://arxiv.org/abs/2401.05566v3.

Vedang Lad, Wes Gurnee, and Max Tegmark. The Remarkable Robustness of LLMs: Stages of Inference?, June 2024. URL http://arxiv.org/abs/2406.19384. arXiv:2406.19384 [cs].

B. A. Levinstein and Daniel A. Herrmann. Still No Lie Detector for Language Models: Probing Empirical and Conceptual Roadblocks, June 2023. URL http://arxiv.org/abs/2307.00175. arXiv:2307.00175 [cs].

Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring How Models Mimic Human Falsehoods, May 2022. URL http://arxiv.org/abs/2109.07958. arXiv:2109.07958 [cs].

Samuel Marks and Max Tegmark. The Geometry of Truth: Emergent Linear Structure in Large Language Model Representations of True/False Datasets, August 2024. URL http://arxiv.org/abs/2310.06824. arXiv:2310.06824 [cs].

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. On Faithfulness and Factuality in Abstractive Summarization. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 1906–1919, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.173. URL https://aclanthology.org/2020.acl-main.173.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and Editing Factual Associations in GPT, January 2023. URL http://arxiv.org/abs/2202.05262. arXiv:2202.05262 [cs].

Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. PROGRESS MEASURES FOR GROKKING VIA MECHANISTIC INTERPRETABILITY. 2023a.

Neel Nanda, Andrew Lee, and Martin Wattenberg. Emergent Linear Representations in World Models of Self-Supervised Sequence Models, September 2023b. URL http://arxiv.org/abs/2309.00941. arXiv:2309.00941 [cs].

Lorenzo Pacchiardi, Alex J. Chan, Sören Mindermann, Ilan Moscovitz, Alexa Y. Pan, Yarin Gal, Owain Evans, and Jan Brauner. How to Catch an AI Liar: Lie Detection in Black-Box LLMs by Asking Unrelated Questions, September 2023. URL http://arxiv.org/abs/2309.15840. arXiv:2309.15840 [cs].

Peter S. Park, Simon Goldstein, Aidan O'Gara, Michael Chen, and Dan Hendrycks. AI Deception: A Survey of Examples, Risks, and Potential Solutions, August 2023. URL http://arxiv.org/abs/2308.14752. arXiv:2308.14752 [cs].

Ethan Perez, Sam Ringer, Kamilė Lukošiūtė, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, Andy Jones, Anna Chen, Ben Mann, Brian Israel, Bryan Seethor, Cameron McKinnon, Christopher Olah, Da Yan, Daniela Amodei, Dario Amodei, Dawn Drain, Dustin Li, Eli Tran-Johnson, Guro Khundadze, Jackson Kernion, James Landis, Jamie Kerr, Jared Mueller, Jeeyoon Hyun, Joshua Landau, Kamal Ndousse, Landon Goldberg, Liane Lovitt, Martin Lucas, Michael Sellitto, Miranda Zhang, Neerav Kingsland, Nelson Elhage, Nicholas Joseph, Noemí Mercado, Nova DasSarma, Oliver Rausch, Robin Larson, Sam McCandlish, Scott Johnston, Shauna Kravec, Sheer El Showk, Tamera Lanham, Timothy Telleen-Lawton, Tom Brown, Tom Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Jack Clark, Samuel R. Bowman, Amanda Askell, Roger Grosse, Danny Hernandez, Deep Ganguli, Evan Hubinger, Nicholas Schiefer, and Jared Kaplan. Discovering Language Model Behaviors with Model-Written Evaluations, December 2022. URL http://arxiv.org/abs/2212.09251. arXiv:2212.09251 [cs].

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. Steering Llama 2 via Contrastive Activation Addition, March 2024. URL http://arxiv.org/abs/2312.06681. arXiv:2312.06681 [cs].

Jeremy Scheurer, Mikita Balesni, and Marius Hobbhahn. LARGE LANGUAGE MODELS CAN STRATEGICALLY DECEIVE THEIR USERS WHEN PUT UNDER PRESSURE. 2024.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Clément Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Christian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, Justin Mao-Jones, Katherine Lee, Kathy Yu, Katie Millican, Lars Lowe Sjoesund, Lisa Lee, Lucas Dixon, Machel Reid, Maciej Mikuła, Mateo Wirth, Michael Sharman, Nikolai Chinaev, Nithum Thain, Olivier Bachem, Oscar Chang, Oscar Wahltinez, Paige Bailey, Paul Michel, Petko Yotov, Rahma Chaabouni, Ramona Comanescu, Reena Jana, Rohan Anil, Ross McIlroy, Ruibo Liu, Ryan Mullins, Samuel L. Smith, Sebastian Borgeaud, Sertan Girgin, Sholto Douglas, Shree Pandya, Siamak Shakeri, Soham De, Ted Klimenko, Tom Hennigan, Vlad Feinberg, Wojciech Stokowiec, Yu-hui Chen, Zafarali Ahmed, Zhitao Gong, Tris Warkentin, Ludovic Peran, Minh Giang, Clément Farabet, Oriol Vinyals, Jeff Dean, Koray Kavukcuoglu, Demis Hassabis, Zoubin Ghahramani, Douglas Eck, Joelle Barral, Fernando Pereira, Eli Collins, Armand Joulin, Noah Fiedel, Evan Senter, Alek Andreev, and Kathleen Kenealy. Gemma: Open Models Based on Gemini Research and Technology, April 2024a. URL http://arxiv.org/abs/2403.08295. arXiv:2403.08295 [cs].

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stanczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Olivier Bachem, Alanna Walton, Aliaksei Severyn, Alicia Parrish, Aliya Ahmad, Allen Hutchison, Alvin Abdagic, Amanda Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony Laforge, Antonia Paterson, Ben Bastian, Bilal Piot, Bo Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry, Chris Welty, Christopher A. Choquette-Choo, Danila Sinopalnikov, David Weinberger, Dimple Vijaykumar, Dominika Rogozińska, Dustin Herbison, Elisa Bandy, Emma Wang, Eric Noland, Erica Moreira, Evan Senter, Evgenii Eltyshev, Francesco Visin, Gabriel Rasskin, Gary Wei, Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna Klimczak-Plucińska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Peng Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost van Amersfoort, Josh Gordon, Josh Lipschultz, Josh Newlan, Ju-yeong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjoesund, Lauren Usui, Laurent Sifre, Lena Heuermann, Leticia Lago, Lilly McNealus, Livio Baldini Soares, Logan Kilpatrick, Lucas Dixon, Luciano Martins, Machel Reid, Manvinder Singh, Mark Iverson, Martin Görner, Mat Velloso, Mateo Wirth, Matt Davidow, Matt Miller, Matthew Rahtz, Matthew Watson, Meg Risdal, Mehran Kazemi, Michael Moynihan, Ming Zhang, Minsuk Kahng, Minwoo Park, Mofi Rahman, Mohit Khatwani, Natalie Dao, Nenshad Bardoliwalla, Nesh Devanathan, Neta Dumai, Nilay Chauhan, Oscar Wahltinez, Pankil Botarda, Parker Barnes, Paul Barham, Paul Michel, Pengchong Jin, Petko Georgiev, Phil Culliton, Pradeep Kuppala, Ramona Comanescu, Ramona Merhej, Reena Jana, Reza Ardeshir Rokni, Rishabh Agarwal, Ryan Mullins, Samaneh Saadat, Sara Mc Carthy, Sarah Perrin, Sébastien M. R. Arnold, Sebastian Krause, Shengyang Dai, Shruti Garg, Shruti Sheth, Sue Ronstrom, Susan Chan, Timothy Jordan, Ting Yu, Tom Eccles, Tom Hennigan, Tomas Kocisky, Tulsee Doshi, Vihan Jain, Vikas Yadav, Vilobh Meshram, Vishal Dharmadhikari, Warren Barkley, Wei Wei, Wenming Ye, Woohyun Han, Woosuk Kwon, Xiang Xu, Zhe Shen, Zhitao Gong, Zichuan Wei, Victor Cotruta, Phoebe Kirk, Anand Rao, Minh Giang, Ludovic Peran, Tris Warkentin, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, D. Sculley, Jeanine Banks, Anca Dragan, Slav Petrov, Oriol

Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Sebastian Borgeaud, Noah Fiedel, Armand Joulin, Kathleen Kenealy, Robert Dadashi, and Alek Andreev. Gemma 2: Improving Open Language Models at a Practical Size, July 2024b. URL https://arxiv.org/abs/2408.00118v2.

Curt Tigges, Oskar John Hollinsworth, Atticus Geiger, and Neel Nanda. Linear Representations of Sentiment in Large Language Models, October 2023. URL https://arxiv.org/abs/2310.15154v1.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open Foundation and Fine-Tuned Chat Models, July 2023. URL https://arxiv.org/abs/2307.09288v2.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Activation Addition: Steering Language Models Without Optimization, June 2024. URL http://arxiv.org/abs/2308.10248. arXiv:2308.10248 [cs].

Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the Wild: a Circuit for Indirect Object Identification in GPT-2 Small. September 2022. URL https://openreview.net/forum?id=NpsVSN6o4ul.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. Qwen2 Technical Report, July 2024. URL https://arxiv.org/abs/2407.10671v4.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. Representation Engineering: A Top-Down Approach to AI Transparency, October 2023. URL https://arxiv.org/abs/2310.01405v3.