Mitigating Content Effects on Reasoning in Language Models Through Fine-Grained Activation Steering

Marco Valentino[1], **Geonhee Kim**[2], **Dhairya Dalal**[3], **Zhixue Zhao**[1], **André Freitas**[2,4,5]
[1]University of Sheffield, UK
[2]University of Manchester, UK
[3]University of Galway, Ireland
[4]Idiap Research Institute, Switzerland
[5]National Biomarker Centre, CRUK-MI, UK
m.valentino@sheffield.ac.uk

Abstract

Large language models (LLMs) exhibit reasoning biases, often conflating content plausibility with formal logical validity. This can lead to wrong inferences in critical domains, where plausible arguments are incorrectly deemed logically valid or vice versa. This paper investigates how content biases on reasoning can be mitigated through activation steering, an inference-time technique that modulates internal activations. Specifically, after localising the layers responsible for formal and plausible inference, we investigate activation steering on a controlled syllogistic reasoning task, designed to disentangle formal validity from content plausibility. An extensive empirical analysis reveals that contrastive steering methods consistently support linear control over content biases. However, a static approach is insufficient to debias all the tested models. We then investigate how to control content effects by dynamically determining the steering parameters through fine-grained conditional methods. By introducing a novel kNN-based conditional approach (`K-CAST`), we demonstrate that conditional steering can effectively reduce biases on unresponsive models, achieving up to 15% absolute improvement in formal reasoning accuracy. Finally, we found that steering for content effects is robust to prompt variations, incurs minimal side effects on multilingual language modeling capabilities, and can partially generalize to different reasoning tasks. In practice, we demonstrate that activation-level interventions offer a scalable inference-time strategy for enhancing the robustness of LLMs, contributing towards more systematic and unbiased reasoning capabilities.

Code, data, and supplementary material — https://github.com/neuro-symbolic-ai/steering_content_effects

1 Introduction

Large language models (LLMs) possess advanced natural and common-sense reasoning capabilities but are prone to content effects, i.e., systematic biases where prior knowledge and believability of content influence logical inference (Bertolazzi, Gatt, and Bernardi 2024; Lampinen et al. 2024). For example, an LLM may incorrectly judge a logically invalid syllogism as valid if its content aligns with common-sense knowledge (e.g., "All students read; some readers are professors; therefore some students are professors"), mirroring human content biases (Lampinen et al. 2024). Such behaviour violates the requirement of formal reasoning, where validity should depend only on logical form, not content. Recent studies have documented these content-based reasoning failures (Bertolazzi, Gatt, and Bernardi 2024), showing that models find factually believable premises easier to "prove" and struggle with abstract or counter-intuitive ones (Lampinen et al. 2024). This undermines LLMs' reliability on formal reasoning, particularly logic-oriented tasks highlighted by Bertolazzi et al. (2024).

On the other hand, prompting strategies alone are insufficient to eliminate content effects. Chain-of-thought (CoT) prompting and related methods (Wei et al. 2022; Kojima et al. 2022) can improve reasoning. However, biases often persist in the generated explanations, and models may still arrive at content-biased conclusions even when "thinking aloud" (Ranaldi, Valentino, and Freitas 2025). In fact, (Bertolazzi et al. 2024) finds that while CoT and fine-tuning boost accuracy on logical deductions, they do not fully remove biases like content believability effects. Similarly, neuro-symbolic approaches have been proposed to improve robustness in formal reasoning with LLMs (Quan et al. 2024, 2025b,a; Pan et al. 2023; Lyu et al. 2023). However, they introduce the complexity of integrating LLMs with external symbolic solvers.

Unlike existing methods, we directly manipulate internal activations to investigate if content-invariance can be effectively enforced through test-time interventions and to gain a deeper understanding of the internal representational mechanisms (see Figure 1). Overall, our contribution and findings can be summarised as follows:

A large-scale dataset to disentangle content from formal reasoning. Expanding on previous work (Bertolazzi, Gatt, and Bernardi 2024; Kim, Valentino, and Freitas 2025; Wysocka et al. 2025), we generate a synthetic dataset leveraging known syllogistic arguments, considering the intersection of plausible/implausible and formally valid/invalid arguments. The dataset includes over 16k arguments generated by instantiating 24 abstract syllogistic schemes with the support of Wordnet (Miller 1995).

Localizing formal and plausible inference. We perform an observational study through probing (Ferreira et al. 2021; Belinkov 2022) to localise information about the validity

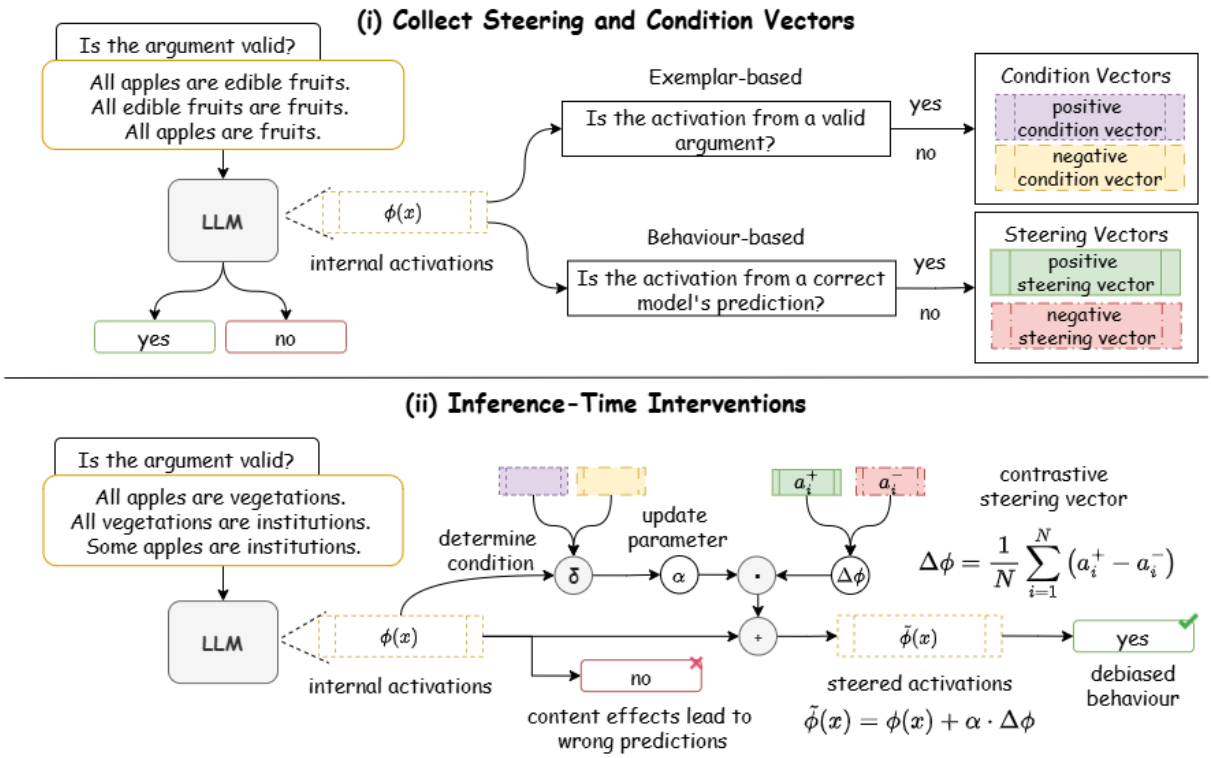


Figure 1: Overview of our methodology for mitigating content effects on reasoning via activation steering. We first curate a controlled syllogistic reasoning dataset designed to disentangle formal validity from content plausibility. Subsequently, after localising the layers mostly responsible for formal and plausible inference through probing, we investigate static and conditional steering methods to debias models' behaviour.

and plausibility of arguments within the models. The experiments reveal that the information is maximally localised in later layers, peaking at the third quarter of the layers in the residual stream across different LLMs.

Evaluating static contrastive steering methods. Leveraging the observational study, we investigate static and contrastive activation steering methods (Panickssery et al. 2023). In general, we found that contrastive steering is effective on most of the tested models. In particular, the experiments reveal that steering vectors can explicitly control models' output along a linear direction depending on the steering parameters, influencing the accuracy on both valid and invalid arguments. However, we found that static steering cannot improve performance on all the tested models.

Adapting and introducing fine-grained conditional steering methods. We adapt the recently proposed conditional activation steering (`CAST`) method (Lee et al. 2025) for content effects, and propose a new fine-grained variation employing a k-NN classifier to dynamically determine the steering parameters (`K-CAST`). We found that such methods can reduce biases on models that are unresponsive to static steering while, at the same time, increasing overall accuracy by up to $\approx 15\%$ absolute value.

Robustness analysis and out-of-distribution generalisation. We investigate the impact of steering for content effects on multilingual language modeling capabilities (Raffel et al. 2020) and out-of-distribution reasoning tasks (Saparov and He 2023; Chan, Gaizauskas, and Zhao 2024). We found that steering is well-localized, incurring minimal side effects on language modeling capabilities. At the same time, we found that steering vectors computed on the synthetic data can generalize to some extent to different reasoning tasks, with some variations across models. These results highlight both the potential of steering to improve targeted reasoning capabilities as well as the persisting challenges in enabling full generalization.

2 Background

Recent research has demonstrated that LLMs exhibit *content effects* in formal reasoning tasks, mirroring cognitive biases that may align or differ from those observed in humans (Dasgupta et al. 2022; Kim, Valentino, and Freitas 2025; Mondorf and Plank 2024; Seals and Shalin 2024; Wysocka et al. 2025; Eisape et al. 2024). These effects arise when the semantic plausibility of a prompt influences the model's reasoning process, often leading to correct conclusions for plausible statements and systematic errors for implausible but logically valid ones (Bertolazzi, Gatt, and Bernardi 2024).

(Dasgupta et al. 2022) first showed that LLMs perform better on reasoning tasks when the content of the problem aligns with world knowledge. In their experiments, models

were significantly more accurate on syllogistic tasks when the conclusions were semantically plausible, even when this plausibility conflicted with the actual logical validity of the argument. This indicates a bias toward material reasoning (reasoning grounded in semantic associations) rather than formal reasoning (reasoning based strictly on logic).

Further work by (Bertolazzi, Gatt, and Bernardi 2024) systematically evaluated LLMs on a broad suite of syllogisms and found that performance dropped sharply for arguments that contradicted commonsense knowledge. This reliance on content plausibility suggests that LLMs are susceptible to semantic interference, failing to uphold the norms of formal logic when they conflict with prior knowledge. (Seals et al. 2023) and (Agarwal, Lyu, and Zhang 2024) also emphasized this discrepancy, showing that even the most capable frontier models tend to conflate logical validity with plausibility. To the best of our knowledge, this is the first work investigating how content effects can be reduced through activation steering techniques (Subramani, Suresh, and Peters 2022; Hernandez, Li, and Andreas 2024; Zou et al. 2023; Turner et al. 2023; Li et al. 2023; Zhao et al. 2024).

3 Methodology

Our goal is to investigate and mitigate *content effects* in LLMs, i.e., systematic biases where semantic plausibility influences logical reasoning. To this end, we design a controlled syllogistic reasoning task, leveraging 24 abstract syllogistic schemes automatically instantiated through taxonomic knowledge from external knowledge bases (Sec. 3.1) and apply activation-level steering techniques to modulate model behavior toward formal validity assessment (Sec. 3.2). Further, we identify the limitations of current state-of-the-art steering methods, and propose a more fine-grained steering approach (Sec. 3.2). Figure 1 provides a high-level overview of the methodology.

3.1 Formal Syllogistic Reasoning

Inspired by recent work (Lample and Charton 2019; Bertolazzi, Gatt, and Bernardi 2024; Wysocka et al. 2025; Kim, Valentino, and Freitas 2025), we evaluate formal reasoning in LLMs through syllogistic arguments. We formalise the task of syllogistic reasoning as a binary classification problem, where the objective is to determine the *formal validity* of a syllogism, whether its conclusion follows logically from its premises, irrespective of the plausibility of the content. Specifically, the model is expected to predict `VALID` or `INVALID` based solely on the logical form in the syllogism structure $\mathcal{S}$.

A syllogism $\mathcal{S}$ is defined as a triple $\mathcal{S} = (P_1, P_2, C)$ where P_1 and P_2 are the two categorical premises, and C is the conclusion. Each statement is expressed in natural language and conforms to standard syllogistic forms (e.g., universal affirmative, universal negative, particular affirmative).

Controlling plausibility for content effect evaluation. To isolate formal validity from world knowledge, we design the task to include the following types of syllogistic arguments:

- **Plausible Valid:** All apples are edible fruits. All edible fruits are fruits. All apples are fruits.
- **Implausible Valid:** All apples are vegetations. All vegetations are institutions. Some apples are institutions.
- **Plausible Invalid:** All apples are edible fruits. All apples are fruits. All edible fruits are fruits.
- **Implausible Invalid:** All apples are institutions. All vegetations are institutions. All apples are vegetations.

This setup allows us to decouple reasoning based on logical form from reasoning based on material content. Models demonstrating robust formal reasoning will maintain consistent accuracy across plausible and implausible conditions by focusing exclusively on argument structure.

Syllogistic arguments generation. We construct a dataset of approximately 16,000 syllogistic arguments in English to systematically analyze content effects. Each argument instantiates one of the 24 formal syllogistic schemas (details in the supplementary material) and is explicitly varied along dimensions of *formal validity* and *semantic plausibility*.

The data generation process begins with the formalization of syllogistic structures in first-order logic (FOL), following prior work on logical reasoning datasets (Bertolazzi, Gatt, and Bernardi 2024; Wysocka et al. 2025). Each logical schema is then converted into natural language templates such as: *All A are B, All B are C, All A are C*.

To control semantic content, we instantiate these syllogistic templates with concrete noun phrases drawn from WordNet[1](Miller 1995) based on taxonomic hierarchies, using hypernym-hyponym relations between concepts.

3.2 Activation Steering

Activation steering is a causal intervention technique for modulating the internal computation of LLMs by linearly modifying hidden activations, also known as activation engineering (Subramani, Suresh, and Peters 2022; Hernandez, Li, and Andreas 2024; Zou et al. 2023; Turner et al. 2023; Li et al. 2023; Zhao et al. 2024). In this work, we adopt both static and conditional activation steering methods.

Contrastive Activation Addition (CAA) computes the steering vector using a set of labelled examples, based on the observed model behavior (Panickssery et al. 2023).

Let $\phi(x) \in \mathbb{R}^d$ denote the activation vector at a chosen layer and token position for input x. Given a set of N contrastive pairs of activations $\mathcal{P} = \{(a_i^+, a_i^-)\}_{i=1}^N$, where $a_i^+ = \phi(x_i^+)$ and $a_i^- = \phi(x_i^-)$ are positive and negative activation vectors derived from inputs x_i^+ and x_i^- leading to desired and contrasting behaviors respectively. The resulting steering vector is the mean difference between positive and negative activations:

$$\Delta\phi = \frac{1}{N}\sum_{i=1}^{N}\left(a_i^+ - a_i^-\right) \tag{1}$$

At inference time, given a new input x, the model is steered by modifying its internal activations $\tilde{\phi}(x) = \phi(x) +$

[1]https://wordnet.princeton.edu/

$\alpha \cdot \Delta\phi$, where α is a scaling hyperparameter determining the strength of the intervention.

We apply CAA to reduce *content effects* by steering activations toward representations associated with content-invariant outputs. To this end, the positive vectors are collected from inputs leading to correct formal validity predictions, while the negative vectors are collected from incorrect predictions affected by content bias.

Conditional Activation Steering (CAST) is a steering method designed to enable selective modulation of model behaviors by conditionally applying activation steering based on the input context (Lee et al. 2025). Unlike traditional activation steering methods that uniformly apply steering vectors across all inputs, CAST introduces a mechanism to determine, at inference time, whether to apply a steering vector based on the similarity of the current input's activation to predefined condition vectors.

Formally, let $\phi(x) \in \mathbb{R}^d$ denote the activation vector at a specified layer and position for input x. Given a labeled dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$, where $y_i \in \{+1, -1\}$ denotes the presence or absence of a target condition on x_i (i.e., the argument is formally valid), `CAST` computes a condition vector ψ_c based on the average aggregation (or PCA) of individual activations vectors $\phi(x_i)$ such that $y_i = +1$. During inference, for a given input x, the similarity between $\phi(x)$ and $\pi_{\psi_c}(\phi(x))$ – i.e., the projection of $\phi(x)$ onto ψ_c – is computed:

$$\text{sim}(\phi(x), \pi_{\psi_c}(\phi(x))) = \frac{\phi(x) \cdot \pi_{\psi_c}(\phi(x))}{\|\phi(x)\| \|\pi_{\psi_c}(\phi(x))\|} \tag{2}$$

In the standard `CAST` method, if the similarity exceeds a predefined threshold θ_c, a corresponding steering vector $\Delta\phi_c$ is applied $\tilde{\phi}(x) = \phi(x) + \alpha \cdot \Delta\phi_c$, where α is a scaling parameter controlling the strength of the intervention.

In this work, we adapt `CAST` to dynamically determine the value of the scaling parameter α, since our empirical analysis on static steering reveals that the sign of α enables explicit control over the accuracy on valid and invalid arguments. In particular, given two condition vectors ψ_{c_+} and ψ_{c_-}, the first computed for *valid* arguments and the second computed for *invalid* arguments, we modify the value of α dynamically according to the following function:

$$f(\alpha, \phi(x), \psi_{c_+}, \psi_{c_-}) = \begin{cases} -\alpha & \text{if } \text{sim}(\phi(x), \pi(\phi(x)_{\psi_{c_+}})) > \text{sim}(\phi(x), \pi(\phi(x)_{\psi_{c_-}})) \\ \alpha & \text{otherwise} \end{cases} \tag{3}$$

Therefore, we perform conditional steering via $\tilde{\phi}(x) = \phi(x) + f(\alpha, \phi(x), \psi_{c_+}, \psi_{c_-}) \cdot \Delta\phi$, where $\Delta\phi$ is a standard contrastive steering vector.

`K-CAST`: kNN-Based Conditional Activation Steering One limitation of `CAST` is that the condition vectors ψ_c are typically computed via aggregating individual activations from different training examples. This can cause a loss of information that undermines the ability to effectively determine the correct condition for test-time intervention. To address this, we introduce an extension to `CAST` that employs a k-Nearest Neighbors (kNN) approach for condition determination, thereby mitigating potential information loss from coarse-grained aggregation methods.

Given a labeled dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$, where $y_i \in \{+1, -1\}$ denotes the presence or absence of a target condition on x_i (i.e., the argument is formally valid), we proceed as follows:

1. For each input x_i in $\mathcal{D}$, compute and store an individual condition activation vector $\psi(x_i)_{y_i}$.
2. At inference time, for a new input x, compute its activation vector $\phi(x)$.
3. Identify the set $\mathcal{N}_k(x) \subset \mathcal{D}$ of k nearest neighbors to $\phi(x)$ based on cosine similarity.
4. Determine the majority condition label $\hat{y}(x)$ among the neighbors:

$$\hat{y}(x) = \text{sign}\left(\sum_{(x_j, y_j) \in \mathcal{N}_k(x)} y_j \right) \tag{4}$$

5. Dynamically determining the steering parameters based on the majority condition label:

$$\tilde{\phi}(x) = \phi(x) + f(\alpha, \Delta\phi, \hat{y}(x)) \tag{5}$$

where $\Delta\phi$ is a standard contrastive steering vector. While this method can be used to arbitrarily adapt the steering method, similarly to `CAST`, we employ it to dynamically determine the value of α at test time via $\tilde{\phi}(x) = \phi(x) - \hat{y}(x) \cdot \alpha \cdot \Delta\phi_c$.

Compared to `CAST`, `K-CAST` allows for a more granular determination of how to apply the steering interventions, leveraging the local structure of the activation space in the training set.

4 Empirical Evaluation

Models. We evaluate the steering performance on three model families, covering different spans of model sizes: Llama (3.2-1b-it, 3.2-3b-it, 3.1-8b) (Grattafiori et al. 2024), Gemma-2 (2b-it, Gemma-2-9b-it) (Team et al. 2024), Qwen 2.5 (1.5b-it, 3b-it, 7b-it) (Bai et al. 2023). We use the instruction-tuned version of each model and evaluate both performance in zero-shot setting and in-context learning (ICL) via few-shot prompts, providing a total of 4 random examples from the training set.

Probing for content effects. To inform subsequent test-time interventions and steering, we performed a preliminary observational study through linear probing (Belinkov 2022; Ferreira et al. 2021) to identify where information about the validity and plausibility of arguments might be encoded within the models. To this end, we employ a linear layer on top of the models' frozen activations after processing a syllogistic argument, classifying whether the argument is valid/invalid and plausible/implausible. Overall, the probing experiments reveal that the information for validity and plausibility is maximally localised in later layers in the residual stream, consistently peaking at the third quarter of the layers

Model	Size	Base Model ($\alpha = 0$)			Steered Model (α_{best})			α_{best}	$\Delta_{abs}^{Acc/CE}$	$\Delta_{rel}^{Acc/CE}$
		Acc ↑	CE ↓	Acc/CE ↑	Acc ↑	CE ↓	Acc/CE ↑			
Zero-shot										
Llama 3.2	1b	58.17	44.04	1.32	73.56	**6.35**	**11.58**	-0.9	**10.26**	**777.27**
	3b	77.79	**17.50**	**4.45**	77.79	17.50	4.45	0.0	0.00	0.00
Llama 3.1	8b	**78.27**	30.77	2.54	**85.10**	14.04	6.06	0.9	3.52	138.58
Gemma 2	2b	73.27	32.43	2.26	74.13	20.83	3.56	1.8	1.30	57.52
	9b	**85.00**	**8.46**	**10.05**	**83.27**	**1.92**	**43.37**	0.6	**33.32**	**331.54**
Qwen 2.5	1.5b	75.67	14.42	5.25	77.79	12.88	6.04	0.3	0.79	15.05
	3b	85.29	7.12	**11.99**	85.29	7.12	11.99	0.0	0.00	0.00
	7b	**88.85**	**5.39**	16.48	**89.90**	**0.96**	**93.65**	-1.5	**77.17**	**468.26**
Few-shot										
Llama 3.2	1b	57.21	45.70	1.25	66.44	**6.08**	**4.94**	-1.5	**3.69**	**295.20**
	3b	**72.79**	28.14	**2.59**	**78.27**	22.31	3.51	0.3	0.92	35.52
Llama 3.1	8b	40.58	**20.26**	2.00	30.67	14.81	2.07	0.3	0.07	3.50
Gemma 2	2b	69.42	**14.23**	4.88	70.00	12.31	5.69	-0.3	0.81	16.60
	9b	**84.61**	15.25	**5.54**	**80.38**	**3.33**	**24.14**	0.3	**18.60**	**335.74**
Qwen 2.5	1.5b	51.92	65.12	0.80	72.69	32.18	2.26	-1.2	1.46	182.50
	3b	86.44	13.91	6.21	86.63	**4.80**	18.02	0.6	**11.81**	**190.18**
	7b	**89.80**	**9.36**	**9.60**	**89.90**	4.81	**18.70**	0.9	9.10	94.79

Table 1: Results of contrastive steering with static values of α. The table compares the performance of unsteered base models ($\alpha = 0$) against the optimal steered performance (α_{best}) with values of α selected from the interval $[-3.0, 3.0]$. Acc/CE is the composite metric (Accuracy/Content Effect). The final columns quantify the Acc/CE gain: Δ_{abs} is the raw increase, and Δ_{rel} is the percentage increase relative to the baseline. The results reveal that contrastive steering is highly effective for most models, except Llama 3.2 3b and Qwen 2.5 3b.

across different LLMs (the detailed results for probing formal validity and plausibility can be found in the supplementary material). Therefore, following the insights from probing, subsequent steering methods intervene within the third quarter of layers at the residual stream corresponding to the last input token position.

Evaluation metrics. We adopt different evaluation metrics to compute the effect of steering on syllogistic reasoning. First, we compute the accuracy (ACC) of each model when assessing the validity of the syllogistic arguments in the test set. In addition, we measure the content effect (CE) based on the difference in accuracy on different subsets of the test set. In particular, we measure both the cross-plausibility CE as the difference in overall accuracy between plausible and implausible arguments, as well as the intra-plausibility CE as the difference in accuracy between valid and invalid arguments given a fixed plausibility value. The overall CE reported in the experiments is computed as the average of cross and intra-plausibility CE. Finally, we report the Acc/CE ratio, as the objective is to obtain maximal accuracy on formal reasoning with minimal content effect.

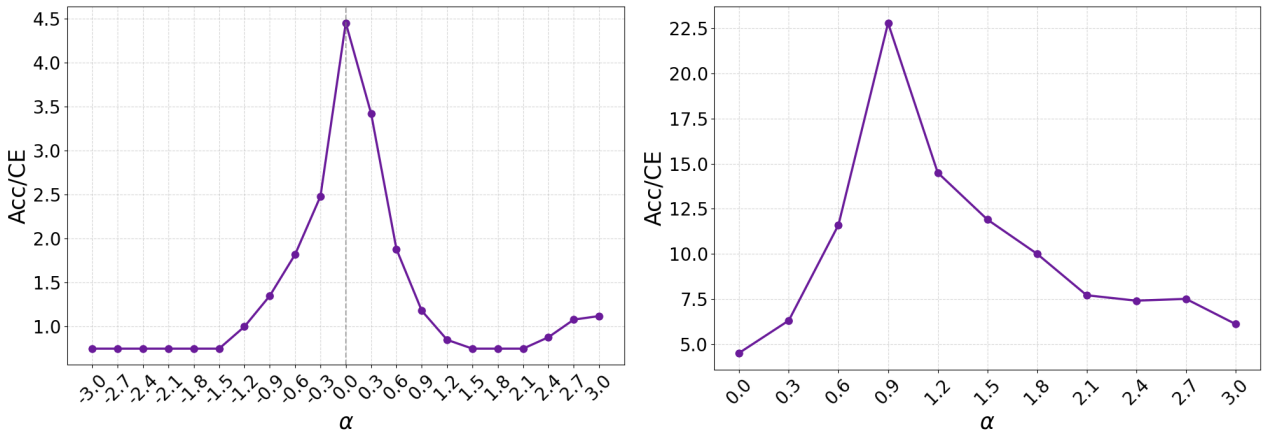
Computing steering vectors. We compute the steering vectors following the methodology described in Section 3.2. In particular, we run each model on a training set composed of 2400 examples equally split across different types of arguments, and select as positive steering vectors the average of the activations that lead to correct predictions, and as negative steering vectors, the average of the activations that lead to wrong predictions.

Model	Size	Acc ↑	CE ↓	Acc/CE ↑	$\Delta_{rel}^{Acc/CE}$
Llama 3.2	3b	77.79	17.50	4.45	-
Qwen 2.5	3b	**85.29**	**7.12**	**11.99**	-
`CAST`					
Llama 3.2	3b	81.04	15.74	5.21	17.07
Qwen 2.5	3b	**85.86**	**4.42**	**19.41**	**61.88**
`K-CAST`					
Llama 3.2	3b	**92.60**	**4.04**	**22.92**	**415.05**
Qwen 2.5	3b	85.28	5.19	16.42	36.94

Table 2: Results of conditional steering on models that are unresponsive to static contrastive steering – i.e, Llama 3.2 3b and Qwen 2.5 7b. We found that both `CAST` and `K-CAST` effectively improve Acc/CE.

4.1 Contrastive Activation Steering

We perform experiments on a test set of 2400 examples equally distributed across different syllogistic schemes. We investigate the effect of steering by varying the value of α between -3 and 3. The results are reported in Table 1.



(a) Llama 3b (Static)

(b) Llama 3b (K−CAST)

Figure 2: (a) Results of steering with static values of α on Llama 3b (note that $\alpha = 0$ represents the performance of the base model without steering). (b) Impact of conditional steering (K−CAST) on Llama 3b. K−CAST increases accuracy while reducing content effects on models that are unresponsive to a static steering approach.

Effectiveness of contrastive steering. The results reveal that contrastive steering is effective for improving Acc/CE on most of the tested models in both zero-shot and ICL settings. Notably, contrastive steering has the highest impact on Llama 3.2 1b with a relative improvement of Acc/CE of up to 777.27%. A substantial improvement can be observed across different families and sizes of models (in particular for Gemma 2 9b and Qwen 2.5 7 via zero-shot). Moreover, we found that for most models, steering for content effect not only improves CE, but also contributes to significant improvements in accuracy on the syllogistic reasoning task (e.g., from 58.17% to 73.56% with Llama 1b). At the same time, steering with a static value of α is ineffective on two zero-shot models – i.e., Llama 3.2 3b and Qwen 2.5 3b.

Steering outperforms ICL. In general, we observe that ICL via few-shot examples is not sufficient to mitigate content effect biases and, in most cases, can have the detrimental effect of reducing accuracy. Contrastive steering, on the other hand, seems to be a much more effective methodology to mitigate reasoning biases in LLMs. This is confirmed by the fact that the best results on syllogistic reasoning are achieved by steered models.

The scaling parameter α enables explicit steering control. In order to investigate why steering is not effective on Llama 3.2 3b and Qwen 2.5 3b, we study the detailed dynamics emerging with different values of α. Here, we observe that, despite being ineffective on some models, contrastive steering can be used to explicitly control the accuracy achieved on valid and invalid arguments by varying the sign of α. Specifically, the results show that setting $\alpha < 0$ generally improves accuracy on *valid* arguments, while $\alpha > 0$ improves accuracy on *invalid* arguments. This observation motivates us to explore conditional steering techniques to dynamically determine the value of α and attempt to steer unresponsive models.

4.2 Conditional Activation Steering

Computing Condition Vectors. Motivated by the observation that the sign of α enables explicit control, we compute condition vectors to identify whether a given model is processing a valid or an invalid argument from the internal activations and then modulate the parameter α accordingly (i.e., setting $\alpha < 0$ if *valid*, and $\alpha > 0$ otherwise). To this end, we collect condition activation vectors for validity from the training set and experiment with both CAST and K−CAST.

Conditional steering is effective on unresponsive models. We found that both CAST and K−CAST are effective in improving Acc/CE for both Llama 3b and Qwen 3b (see Table 2). Moreover, while the results on Qwen show that CAST and K−CAST have a similar effect, the results on Llama reveal that K−CAST is significantly more effective, leading to an absolute increase in accuracy of up to 15%. Figure 2 (b) shows the impact of K−CAST on Llama 3b with different values of α (with the sign dynamically determined).

4.3 Robustness to Prompt Perturbations

To test the robustness of steering performance to prompt perturbations (Mizrahi et al. 2024), we construct a set of prompt variants by employing instruction templates different from those used in the training set (i.e. instruction template paraphrasing). Following (Mizrahi et al. 2024), we employ two prompting strategies proven effective in prior research: (1) *Instruction template rephrasing*: we use GPT-4.5 to paraphrase a seed instruction template (Lester, Al-Rfou, and Constant 2021; Gonen et al. 2022; Honovich et al. 2023a); (2) *Instruction induction*: inspired by (Honovich et al. 2023b), we provide five input-output pairs and ask GPT-4.5 to generate the possible instructions. Given a set of prompt variations (see supplementary material), we compute the steering vectors using the original prompt and randomly select a variant at inference time.

Model	Size	English			Chinese			German		
		$\text{PPL}_{\alpha=0}\downarrow$	$\text{PPL}_{\alpha_{\text{best}}}\downarrow$	$\Delta\%$	$\text{PPL}_{\alpha=0}\downarrow$	$\text{PPL}_{\alpha_{\text{best}}}\downarrow$	$\Delta\%$	$\text{PPL}_{\alpha=0}\downarrow$	$\text{PPL}_{\alpha_{\text{best}}}\downarrow$	$\Delta\%$
Llama 3.2	1b	24.29	24.77	**1.98**	49.81	51.35	**3.09**	20.16	20.52	**1.79**
Gemma 2	9b	18.95	20.58	8.60	34.00	38.17	12.26	16.04	17.43	8.67
Qwen 2.5	7b	**14.59**	**15.17**	3.98	**18.41**	**19.07**	3.58	**11.18**	**11.58**	3.57

Model	Size	ProntoQA			Rulebreakers		
		$\text{ACC}_{\alpha=0}\uparrow$	$\text{ACC}_{\alpha_{\text{best}}}\uparrow$	$\Delta\%$	$\text{ACC}_{\alpha=0}\uparrow$	$\text{ACC}_{\alpha_{\text{best}}}\uparrow$	$\Delta\%$
Llama 3.2	1b	49.6	53.6	**8.1**	40.2	38.6	-4.0
Gemma 2	9b	**62.2**	52.2	-16.1	**92.0**	85.6	-6.9
Qwen 2.5	7b	53.6	**56.4**	5.2	88.2	**88.2**	**0.0**

Table 3: (Top) Impact of steering on multilingual language modeling capabilities. (Bottom) Generalization to OOD logical reasoning tasks. The results demonstrate that steering for content effects incurs minimal side effects on multilingual language modeling capabilities, and can generalize to some extent to out-of-distribution reasoning tasks.

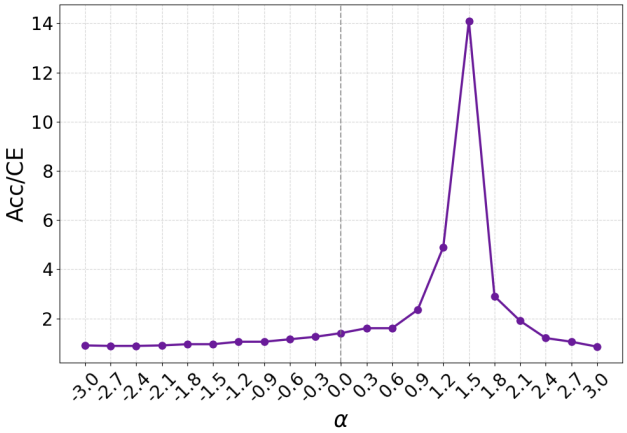


Figure 3: Robustness of steering to prompt variations on Llama 1b (i.e. Acc/CE). The results reveal that, despite specific variations deriving from perturbations applied at test time, the overall effectiveness of steering remains unaltered.

Steering is robust to prompt variations. The results in Figure 3 on Llama 1b (i.e., the model with the best Acc/CE improvement on the original prompt) reveal that, despite specific variation in the values of α_{best} and some noise deriving from prompt variation, the overall effectiveness of steering remains unaltered. A similar trend is also observable for other models (see supplementary material).

4.4 Impact on Non-Target Capabilities

Ideally, the steering effect should be localised – i.e., do not impact non-target capabilities. In this section, we particularly consider the language modeling capability and the reasoning capability on out-of-distribution (OOD) tasks: information-informed reasoning and multi-premise deductive reasoning. For each experiment, we compare the performance between the model without steering ($\alpha = 0$ in Table 3) against the model with steering (α_{best}). If steering is perfectly localized, the two should perform indistinguishably on these non-target tasks.

Multilingual language modeling. We draw 2,000 examples per language from the C4 dataset (Raffel et al. 2020) and compute the average perplexity on causal language modeling over sequences of length 1,024. Table 3 shows that content-effect steering leaves multilingual modeling nearly intact. For example, on English text, Llama 3.2 1b model's perplexity changes only from 24.29 (baseline) to 24.77 (steered). Gemma exhibits the most significant relative increase, but even there, the gap remains small; across all languages and model sizes, perplexity deviations stay within a few percent.

OOD reasoning tasks. We test the steering impact on two reasoning tasks (i.e., ProntoQA (Saparov and He 2023) and Rulebreakers (Chan, Gaizauskas, and Zhao 2024)) that were not presented during the steering modulating process. ProntoQA is a synthetic task designed to test deductive reasoning on multiple natural language premises. Rulebreakers is a task for evaluating LLMs ability to distinguish whether logical entailment diverges from factually acceptable inference. Table 3 show that adopting the steering vectors computed on syllogisms generalise well, especially on ProntoQA with Llama and Qwen (+8.1% and +5.2%), whereas Gemma experiences a substantial performance drop on both tasks, most notably a 16.1% decrease on ProntoQA (this aligns with the higher drops observed on language modeling).

These findings underscore both the promise of steering for enhancing targeted reasoning and the persistent challenge of achieving a complete and robust OOD generalization.

5 Conclusion

This paper investigated how to mitigate content effects in LLMs' reasoning. We systematically tested static and conditional activation steering techniques to disentangle formal validity from content plausibility in syllogistic reasoning. Our results indicate that steering is particularly effective in reducing content effect and improving accuracy in formal reasoning. In general, this paper demonstrates that activation-level interventions offer a scalable inference-time strategy and can contribute towards more systematic and unbiased formal reasoning in LLMs.

Acknowledgments

This work was partially funded by the SNSF project NeuMath (200021_204617), by the CRUK National Biomarker Centre, and supported by the Manchester Experimental Cancer Medicine Centre and the NIHR Manchester Biomedical Research Centre.

References

Agarwal, S.; Lyu, Y.; and Zhang, H. 2024. Mind the Gap: From Plausible to Valid Self-Explanations in Large Language Models. *arXiv preprint arXiv:2405.02706*.

Bai, J.; Bai, S.; Chu, Y.; Cui, Z.; Dang, K.; Deng, X.; Fan, Y.; Ge, W.; Han, Y.; Huang, F.; et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Belinkov, Y. 2022. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1): 207–219.

Bertolazzi, L.; Gatt, A.; and Bernardi, R. 2024. A Systematic Analysis of Large Language Models as Soft Reasoners: The Case of Syllogistic Inferences. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 13882–13905. Miami, Florida, USA: Association for Computational Linguistics.

Bertolazzi, L.; Melloni, A.; Ferrari, M.; and Dragoni, M. 2024. A Systematic Evaluation of Logical Reasoning in Large Language Models. In *International Conference on Learning Representations (ICLR)*.

Chan, J.; Gaizauskas, R.; and Zhao, Z. 2024. Rulebreakers Challenge: Revealing a Blind Spot in Large Language Models' Reasoning with Formal Logic. *arXiv preprint arXiv:2410.16502*.

Dasgupta, I.; Lampinen, A.; Chan, S. C.; Creswell, A.; McClelland, J. L.; and Hill, F. 2022. Content Effects on Logical Reasoning in Transformers. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Eisape, T.; Tessler, M.; Dasgupta, I.; Sha, F.; Steenkiste, S.; and Linzen, T. 2024. A Systematic Comparison of Syllogistic Reasoning in Humans and Language Models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 8418–8437.

Ferreira, D.; Rozanova, J.; Thayaparan, M.; Valentino, M.; and Freitas, A. 2021. Does My Representation Capture X? Probe-Ably. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, 194–201.

Gonen, H.; Iyer, S.; Blevins, T.; Smith, N. A.; and Zettlemoyer, L. 2022. Demystifying prompts in language models via perplexity estimation. *arXiv preprint arXiv:2212.04037*.

Grattafiori, A.; Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Vaughan, A.; et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Hernandez, E.; Li, B. Z.; and Andreas, J. 2024. Inspecting and Editing Knowledge Representations in Language Models. In *First Conference on Language Modeling*.

Honovich, O.; Scialom, T.; Levy, O.; and Schick, T. 2023a. Unnatural Instructions: Tuning Language Models with (Almost) No Human Labor. In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 14409–14428. Toronto, Canada: Association for Computational Linguistics.

Honovich, O.; Shaham, U.; Bowman, S. R.; and Levy, O. 2023b. Instruction Induction: From Few Examples to Natural Language Task Descriptions. In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1935–1952. Toronto, Canada: Association for Computational Linguistics.

Kim, G.; Valentino, M.; and Freitas, A. 2025. Reasoning Circuits in Language Models: A Mechanistic Interpretation of Syllogistic Inference. In Che, W.; Nabende, J.; Shutova, E.; and Pilehvar, M. T., eds., *Findings of the Association for Computational Linguistics: ACL 2025*, 10074–10095. Vienna, Austria: Association for Computational Linguistics. ISBN 979-8-89176-256-5.

Kojima, T.; Gu, S. S.; Reid, M.; Matsuo, Y.; and Iwasawa, Y. 2022. Large Language Models are Zero-Shot Reasoners. *arXiv preprint arXiv:2205.11916*.

Lampinen, A. K.; Dasgupta, I.; Chan, S. C. Y.; Sheahan, H. R.; Creswell, A.; Kumaran, D.; McClelland, J. L.; and Hill, F. 2024. Language models, like humans, show content effects on reasoning tasks. *PNAS Nexus*, 3(7): pgae233.

Lample, G.; and Charton, F. 2019. Deep learning for symbolic mathematics. *arXiv preprint arXiv:1912.01412*.

Lee, B. W.; Padhi, I.; Ramamurthy, K. N.; Miehling, E.; Dognin, P.; Nagireddy, M.; and Dhurandhar, A. 2025. Programming Refusal with Conditional Activation Steering. In *The Thirteenth International Conference on Learning Representations*.

Lester, B.; Al-Rfou, R.; and Constant, N. 2021. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*.

Li, K.; Patel, O.; Viégas, F.; Pfister, H.; and Wattenberg, M. 2023. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems*, 36: 41451–41530.

Lyu, Q.; Havaldar, S.; Stein, A.; Zhang, L.; Rao, D.; Wong, E.; Apidianaki, M.; and Callison-Burch, C. 2023. Faithful Chain-of-Thought Reasoning. In Park, J. C.; Arase, Y.; Hu, B.; Lu, W.; Wijaya, D.; Purwarianti, A.; and Krisnadhi, A. A., eds., *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, 305–329. Nusa Dua, Bali: Association for Computational Linguistics.

Miller, G. A. 1995. WordNet: a lexical database for English. *Communications of the ACM*, 38(11): 39–41.

Mizrahi, M.; Kaplan, G.; Malkin, D.; Dror, R.; Shahaf, D.; and Stanovsky, G. 2024. State of What Art? A Call for Multi-Prompt LLM Evaluation. *Transactions of the Association for Computational Linguistics*, 12: 933–949.

Mondorf, P.; and Plank, B. 2024. Comparing Inferential Strategies of Humans and Large Language Models in Deductive Reasoning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 9370–9402.

Pan, L.; Albalak, A.; Wang, X.; and Wang, W. 2023. Logic-LM: Empowering Large Language Models with Symbolic Solvers for Faithful Logical Reasoning. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Findings of the Association for Computational Linguistics: EMNLP 2023*, 3806–3824. Singapore: Association for Computational Linguistics.

Panickssery, N.; Gabrieli, N.; Schulz, J.; Tong, M.; Hubinger, E.; and Turner, A. M. 2023. Steering llama 2 via contrastive activation addition. *arXiv preprint arXiv:2312.06681*.

Quan, X.; Valentino, M.; Carvalho, D. S.; Dalal, D.; and Freitas, A. 2025a. PEIRCE: Unifying Material and Formal Reasoning via LLM-Driven Neuro-Symbolic Refinement. *arXiv preprint arXiv:2504.04110*.

Quan, X.; Valentino, M.; Dennis, L. A.; and Freitas, A. 2024. Verification and Refinement of Natural Language Explanations through LLM-Symbolic Theorem Proving. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2933–2958. Miami, Florida, USA: Association for Computational Linguistics.

Quan, X.; Valentino, M.; Dennis, L. A.; and Freitas, A. 2025b. Faithful and Robust LLM-Driven Theorem Proving for NLI Explanations. In Che, W.; Nabende, J.; Shutova, E.; and Pilehvar, M. T., eds., *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 17734–17755. Vienna, Austria: Association for Computational Linguistics. ISBN 979-8-89176-251-0.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140): 1–67.

Ranaldi, L.; Valentino, M.; and Freitas, A. 2025. Improving Chain-of-Thought Reasoning via Quasi-Symbolic Abstractions. In Che, W.; Nabende, J.; Shutova, E.; and Pilehvar, M. T., eds., *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 17222–17240. Vienna, Austria: Association for Computational Linguistics. ISBN 979-8-89176-251-0.

Saparov, A.; and He, H. 2023. Language Models Are Greedy Reasoners: A Systematic Formal Analysis of Chain-of-Thought. In *The Eleventh International Conference on Learning Representations*.

Seals, J.; Dasgupta, S.; Kumar, A.; and Ghosh, S. 2023. Do LLMs Exhibit Content Effects? An Investigation of Human-like Biases in Language Models. *CEUR Workshop Proceedings*, 3606: 111–125.

Seals, S.; and Shalin, V. 2024. Evaluating the Deductive Competence of Large Language Models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 8606–8622.

Subramani, N.; Suresh, N.; and Peters, M. E. 2022. Extracting latent steering vectors from pretrained language models. *arXiv preprint arXiv:2205.05124*.

Team, G.; Riviere, M.; Pathak, S.; Sessa, P. G.; Hardin, C.; Bhupatiraju, S.; Hussenot, L.; Mesnard, T.; Shahriari, B.; Ramé, A.; et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Turner, A. M.; Thiergart, L.; Leech, G.; Udell, D.; Vazquez, J. J.; Mini, U.; and MacDiarmid, M. 2023. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35: 24824–24837.

Wysocka, M.; Carvalho, D.; Wysocki, O.; Valentino, M.; and Freitas, A. 2025. SylloBio-NLI: Evaluating Large Language Models on Biomedical Syllogistic Reasoning. In Chiruzzo, L.; Ritter, A.; and Wang, L., eds., *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 7235–7258. Albuquerque, New Mexico: Association for Computational Linguistics. ISBN 979-8-89176-189-6.

Zhao, Y.; Devoto, A.; Hong, G.; Du, X.; Gema, A. P.; Wang, H.; He, X.; Wong, K.-F.; and Minervini, P. 2024. Steering knowledge selection behaviours in LLMs via sae-based representation engineering. *arXiv preprint arXiv:2410.15999*.

Zou, A.; Phan, L.; Chen, S.; Campbell, J.; Guo, P.; Ren, R.; Pan, A.; Yin, X.; Mazeika, M.; Dombrowski, A.-K.; et al. 2023. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*.