

---

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

---

**Marco Valentino<sup>1,2</sup>, Geonhee Kim<sup>3</sup>, Dhairyा Dalal<sup>4</sup>, Zhixue Zhao<sup>1</sup>, André Freitas<sup>2,3,5</sup>**

<sup>1</sup>University of Sheffield, UK

<sup>2</sup>Idiap Research Institute, Switzerland

<sup>3</sup>University of Manchester, UK

<sup>4</sup>University of Galway, Ireland

<sup>5</sup>National Biomarker Centre, CRUK-MI, UK

## Abstract

Large language models (LLMs) frequently demonstrate reasoning limitations, often conflating content plausibility (i.e., material inference) with logical validity (i.e., formal inference). This can result in biased inferences, where plausible arguments are incorrectly deemed logically valid or vice versa. Understanding and mitigating this limitation is critical, as it undermines the trustworthiness and generalizability of LLMs in applications that demand rigorous logical consistency, such as legal reasoning, scientific analysis, and safety-critical decision support. This paper investigates the problem of mitigating content biases on formal reasoning through activation steering, an inference-time intervention technique that directly modulates internal model activations. Specifically, we first curate a controlled syllogistic reasoning dataset covering 24 types of logical argument schemes designed to disentangle formal validity from content plausibility. After localising the layers responsible for formal and material inference through probing, we investigate contrastive activation steering methods for test-time interventions. An extensive empirical analysis on different LLMs (i.e. Llama, Gemma, and Qwen) reveals that contrastive steering consistently supports linear control over content biases. However, we observe that a static steering approach is insufficient for achieving improvements on all the tested models. We then leverage the possibility to control content effects by dynamically determining the value of the steering parameters via fine-grained conditional methods. We found that conditional steering is effective in reducing biases on unresponsive models, achieving up to 15% absolute improvement in formal reasoning accuracy with a newly introduced kNN-based conditional method (K-CAST). Finally, additional experiments reveal that steering for content effects is robust to prompt variations, incurs minimal side effects on multilingual language modeling capabilities, and can partially generalize to out-of-distribution reasoning tasks. Practically, this paper demonstrates that activation-level interventions can offer a scalable inference-time strategy for enhancing the robustness of LLMs, contributing towards more systematic and unbiased formal reasoning<sup>1</sup>.

## 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 [4, 17]. For example, an LLM may incorrectly judge a logically

---

<sup>1</sup>Correspondence email: ac4mv@sheffield.ac.uk. Code and data will be made available upon publication.

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 [17]. Such behavior 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 in LLMs [4], showing that models, like humans, find factually believable premises easier to “prove” and struggle with abstract or counter-intuitive ones [17]. This undermines LLMs reliability on formal reasoning tasks, particularly syllogistic logic-oriented tasks highlighted by [5].

On the other hand, prompting strategies alone are insufficient to eliminate content effects. Chain-of-thought (CoT) prompting [38] and related methods (e.g. zero-shot CoT [16]) can improve reasoning performance by eliciting step-by-step logic. However, biases often persist in these generated explanations, and models may still arrive at content-biased conclusions even when “thinking aloud” [31]. In fact, [5] finds that while CoT-based in-context learning boosts accuracy on logical deductions, it does not fully remove biases like content believability effects. Similarly, straightforward instructions to “ignore prior knowledge” tend to be ineffective because the model’s latent representations already entangle content with reasoning [7]. Supervised fine-tuning on reasoning data can mitigate some biases [5], but it requires extensive data curation and does not explicitly guarantee content neutrality in the reasoning process. Similarly, neuro-symbolic approaches have been proposed to improve robustness in formal reasoning with LLMs [29, 28, 26, 22]. However, they typically introduce the complexity of integrating LLMs with external symbolic solvers through autoformalization [39].

In this paper, we aim to go beyond these baselines by directly intervening internally into the model’s reasoning to enforce content-invariance (see Figure 1). We focus on syllogistic reasoning, a task that enables us to systematically disentangle formal reasoning from content by leveraging well-known logical-deductive argument schemes in the literature. Overall our contribution and findings can be summarised as follows:

**A large-scale dataset to disentangle formal and material inference.** Expanding on previous work [4, 15, 40], we generate a synthetic dataset leveraging known syllogistic arguments considering the intersection of plausible/imausible and formally valid/invalid arguments. The dataset includes over 16k arguments generated by instantiating 24 abstract syllogistic schemes with the support of external knowledge bases (i.e., Wordnet [23]).

**Probing and localizing formal and material inference.** We perform an observational study through probing [9, 3] to localise information about the validity 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 [27]. 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 [19] 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 [30] and out-of-distribution reasoning tasks [32, 6]. 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.

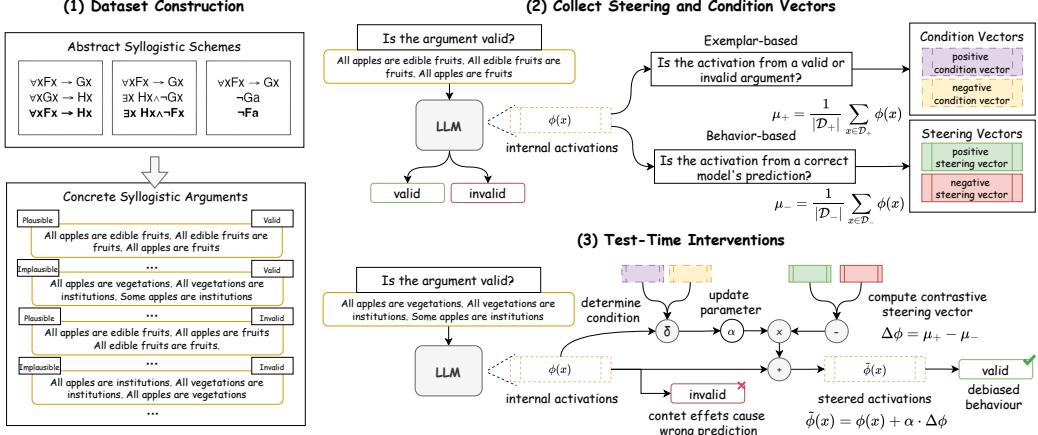


Figure 1: Overview of our methodology for mitigating content effects on formal reasoning via activation steering. We first curate a controlled syllogistic reasoning dataset designed to disentangle formal validity from content plausibility. Subsequently, after localising the layers mostly responsible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

## 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 [7, 15, 25, 34, 40, 8]. 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 [4].

[7] 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 [4] 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. [33] and [1] also emphasized this discrepancy, showing that even the most capable frontier models tend to conflate logical validity with content-driven plausibility. To the best of our knowledge, this is the first work investigating how content effects can be reduced through activation steering techniques [35, 12, 42, 37, 21, 41].

## 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. 3.2.1). Figure 1 provides a high-level overview of the methodology.

### 3.1 Formal Reasoning Task on Syllogistic Arguments

Inspired by recent work [18, 4, 40, 15], 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:

|                            |   |
|----------------------------|---|
| <b>Plausible Valid</b>     | All apples are edible fruits. All edible fruits are fruits. All apples are fruits.          |
| <b>Implausible Valid</b>   | All apples are vegetations. All vegetations are institutions. Some apples are institutions. |
| <b>Plausible Invalid</b>   | All apples are edible fruits. All apples are fruits. All edible fruits are fruits.          |
| <b>Implausible Invalid</b> | 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 (see Appendix), ranging from categorical to disjunctive forms, and is explicitly varied along dimensions of *formal validity* and *semantic plausibility*.

Our data generation process begins with the formalization of syllogistic structures in first-order logic (FOL), following prior work on logical reasoning datasets [4, 40]. For example, an AA1 schema is defined as:

$$\forall x (A(x) \rightarrow B(x)), \quad \forall x (B(x) \rightarrow C(x)) \Rightarrow \forall x (A(x) \rightarrow C(x))$$

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<sup>2</sup>[23] based on its taxonomic hierarchies, using hypernym-hyponym relations.

### 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 [35, 12, 42, 37, 21, 41]. In this work, we adopt both static and conditional activation steering methods.

**Contrastive Activation Steering (CAA)** computes the intervention vector using a set of labeled examples, based on the observed model behavior [27]. Let  $\phi(x) \in \mathbb{R}^d$  denote the activation vector at a chosen layer and token position for input  $x$ . Given a training set  $\mathcal{D} = \{(x_i, y_i)\}$ , where  $y_i \in \{+1, -1\}$  labels whether the model’s output exhibits the intended behavior, we compute average activations  $\mu_+ = \frac{1}{|\mathcal{D}_+|} \sum_{x \in \mathcal{D}_+} \phi(x)$  and  $\mu_- = \frac{1}{|\mathcal{D}_-|} \sum_{x \in \mathcal{D}_-} \phi(x)$ , with  $\mathcal{D}_+ = \{x_i \mid y_i = +1\}$  and  $\mathcal{D}_- = \{x_i \mid y_i = -1\}$ .

The resulting steering direction is  $\Delta\phi = \mu_+ - \mu_-$ . At inference time, we steer the model by modifying its internal activations:

$$\tilde{\phi}(x) = \phi(x) + \alpha \cdot \Delta\phi \tag{1}$$

<sup>2</sup><https://wordnet.princeton.edu/>

where  $\alpha$  is a scaling hyperparameter. This method has proven effective at shifting model outputs in desired directions with minimal degradation to fluency or relevance.

We apply CAA to reduce *content effects* by steering activations toward representations associated with content-invariant outputs. To this end,  $\mu_+$  is computed from activations leading to correct formal validity predictions, while  $\mu_-$  is computed 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 [19]. 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  $\psi_c$  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  $\text{proj}\phi(x)_{\psi_c}$  – i.e., The projection of  $\phi(x)$  onto  $\psi_c$  – is computed, typically using cosine similarity:

$$\text{sim}(\phi(x), \text{proj}\phi(x)_{\psi_c}) = \frac{\phi(x) \cdot \text{proj}\phi(x)_{\psi_c}}{\|\phi(x)\| \|\text{proj}\phi(x)_{\psi_c}\|} \quad (2)$$

In the standard CAST method, if the similarity exceeds a predefined threshold  $\theta_c$ , a corresponding steering vector  $\Delta\phi_c$  is applied  $\tilde{\phi}(x) = \phi(x) + \alpha \cdot \Delta\phi_c$ , where  $\alpha$  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  $\alpha$ , since our empirical analysis on static contrastive steering reveals that the sign of  $\alpha$  enables explicit control over the accuracy on valid and invalid arguments. In particular, given two condition vectors  $\psi_{c+}$  and  $\psi_{c-}$ , the first computed for *valid* arguments and the second computed for *invalid* arguments, we modify the value of  $\alpha$  dynamically according to the following function:

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

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.

### 3.2.1 K-CAST: kNN-Based Conditional Activation Steering

One limitation of CAST is that the condition vectors  $\psi_c$  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:

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

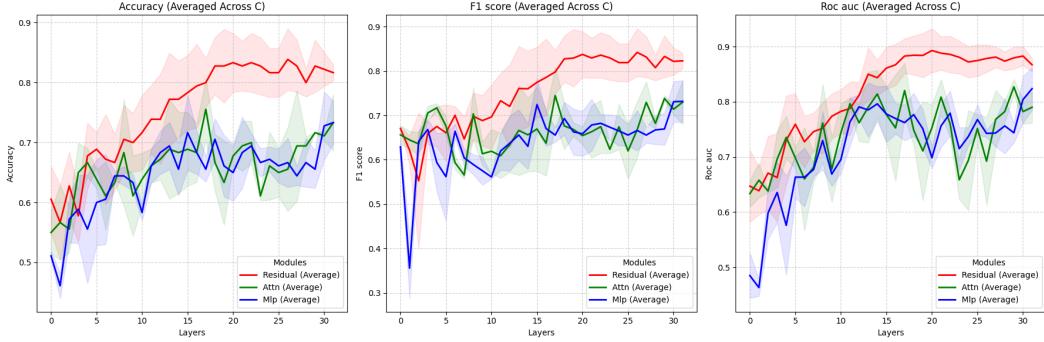


Figure 2: Linear probing results for formal validity on Llama-3.1 8b. The probing experiments reveal that the information for both validity and plausibility is maximally localised in later layers, peaking at the last third quarter of the layers in the residual stream across different LLMs (see Appendix).

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

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

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  $\alpha$  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) [11], Gemma-2 (2b-it, Gemma-2-9b-it) [36], Qwen 2.5 (1.5b-it, 3b-it, 7b-it) [2]. 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 [3, 9] 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 frozen activations of the models after processing a syllogistic argument, classifying whether the argument is valid/invalid and plausible/implausible. Figure 2 shows the results for formal validity on Llama 3.1-8B instruct (the results for plausibility and other models are reported in the Appendix). Overall, the probing experiments reveal that the information for both validity and plausibility is maximally localised in later layers, peaking at the last third quarter of the layers in the residual stream across different LLMs. Therefore, subsequent steering methods intervene on the last third quarter of layers at 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, since the target 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

|                  |      | $\alpha = 0$   |                 |                   | $\alpha \neq 0$ |                 |                   |                 |               |
|------------------|------|----------------|-----------------|-------------------|-----------------|-----------------|-------------------|-----------------|---------------|
| Model            | Size | Acc $\uparrow$ | CE $\downarrow$ | Acc/CE $\uparrow$ | Acc $\uparrow$  | CE $\downarrow$ | Acc/CE $\uparrow$ | $\alpha_{best}$ | $\Delta\%$    |
| <b>Zero-shot</b> |      |                |                 |                   |                 |                 |                   |                 |               |
| Llama 3.2        | 1b   | 58.17          | 44.04           | 1.32              | 73.56           | <b>6.35</b>     | <b>11.58</b>      | -0.9            | <b>777.27</b> |
|                  | 3b   | 77.79          | <b>17.50</b>    | <b>4.45</b>       | 77.79           | 17.50           | 4.45              | 0.0             | 0.00          |
| Llama 3.1        | 8b   | <b>78.27</b>   | 30.77           | 2.54              | <b>85.10</b>    | 14.04           | 6.06              | 0.9             | 138.58        |
| Gemma 2          | 2b   | 73.27          | 32.43           | 2.26              | 74.13           | 20.83           | 3.56              | 1.8             | 57.52         |
|                  | 9b   | <b>85.00</b>   | <b>8.46</b>     | <b>10.05</b>      | <b>83.27</b>    | <b>1.92</b>     | <b>43.37</b>      | 0.6             | <b>331.54</b> |
| Qwen 2.5         | 1.5b | 75.67          | 14.42           | 5.25              | 77.79           | 12.88           | 6.04              | 0.3             | 15.05         |
|                  | 3b   | 85.29          | 7.12            | 11.99             | 85.29           | 7.12            | 11.99             | 0.0             | 0.00          |
|                  | 7b   | <b>88.85</b>   | <b>5.39</b>     | <b>16.48</b>      | <b>89.90</b>    | <b>0.96</b>     | <b>93.65</b>      | -1.5            | <b>468.26</b> |
| <b>ICL</b>       |      |                |                 |                   |                 |                 |                   |                 |               |
| Llama 3.2        | 1b   | 57.21          | 45.70           | 1.25              | 66.44           | 6.08            | <b>4.94</b>       | -1.5            | <b>295.2</b>  |
|                  | 3b   | <b>72.79</b>   | 28.14           | <b>2.59</b>       | <b>78.27</b>    | 22.31           | 3.51              | 0.3             | 35.52         |
| Llama 3.1        | 8b   | 40.58          | <b>20.26</b>    | 2.00              | 30.67           | <b>14.81</b>    | 2.07              | 0.3             | 3.5           |
| Gemma 2          | 2b   | 69.42          | <b>14.23</b>    | 4.88              | 70.00           | 12.31           | 5.69              | -0.3            | 16.6          |
|                  | 9b   | <b>84.61</b>   | 15.25           | <b>5.54</b>       | <b>80.38</b>    | <b>3.33</b>     | <b>24.14</b>      | 0.3             | <b>335.74</b> |
| Qwen 2.5         | 1.5b | 51.92          | 65.12           | 0.80              | 72.69           | 32.18           | 2.26              | -1.2            | 182.5         |
|                  | 3b   | 86.44          | 13.91           | 6.21              | 86.63           | 4.80            | 18.02             | 0.6             | <b>190.18</b> |
|                  | 7b   | <b>89.80</b>   | <b>9.36</b>     | <b>9.60</b>       | <b>89.90</b>    | <b>4.81</b>     | <b>18.70</b>      | 0.9             | 94.79         |

Table 1: Results of contrastive steering with static values of  $\alpha$ . The table shows the results obtained with the best value of  $\alpha$  based on accuracy and content effect ratio (ACC/CE) selected from the interval  $[-3.0, 3.0]$ . A value of  $\alpha_{best} = 0$  indicates that the best results are achieved without steering.  $\Delta\%$  represents the relative improvement of Acc/CE for each model. We found that contrastive steering is highly effective for most models, except Llama 3.2 3b and Qwen 2.5 3b.

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.

#### 4.1 Contrastive Activation Steering

We perform experiments on a test set of 2400 examples equally distributed across different types of arguments and schemes. We investigate the effect of steering varying the value of  $\alpha$  between -3 and 3. The results are reported in Table 1. We summarise the main results and observations below.

**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  $\alpha$  seems to be 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.

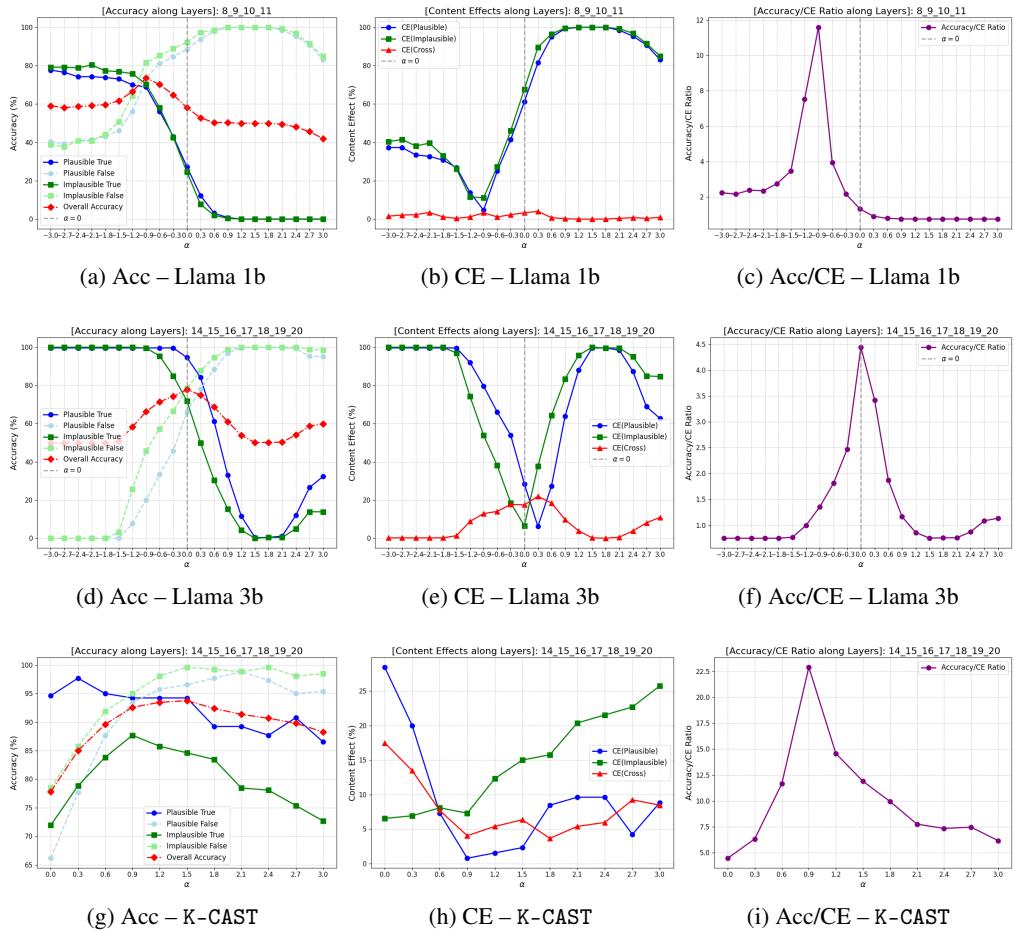


Figure 3: (Top) Example of effective (Llama 1b) and ineffective (Llama 3b) contrastive steering with static values of  $\alpha$ . (Bottom) Impact of conditional activation steering (K-CAST) on Llama 3b. Contrary to static steering, K-CAST leads to a significant increase in accuracy for Llama 3b (i.e., up to 15%) while substantially reducing content effect.

**The scaling parameter  $\alpha$  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  $\alpha$  (Figure 3, top). 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  $\alpha$ . 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  $\alpha$  and attempt to steer unresponsive models.

## 4.2 Conditional Activation Steering

**Computing Condition Vectors.** Motivated by the observation that the sign of  $\alpha$  can enable explicit control on the task, 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  $\alpha$  accordingly (i.e., setting  $\alpha < 0$  if the condition is *valid*, and  $\alpha > 0$  otherwise). To this end, we collect condition activation vectors for validity from the training set (separating valid from invalid arguments) 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

| Model              | Size | Acc $\uparrow$ | CE $\downarrow$ | Acc/CE $\uparrow$ | $\Delta\%$    |
|--------------------|------|----------------|-----------------|-------------------|---------------|
| <b>No Steering</b> |      |                |                 |                   |               |
| Llama 3.2          | 3b   | 77.79          | 17.50           | 4.45              | -             |
| Qwen 2.5           | 3b   | <b>85.29</b>   | <b>7.12</b>     | <b>11.99</b>      | -             |
| <b>CAST</b>        |      |                |                 |                   |               |
| Llama 3.2          | 3b   | 81.04          | 15.74           | 5.21              | 17.07         |
| Qwen 2.5           | 3b   | <b>85.86</b>   | <b>4.42</b>     | <b>19.41</b>      | <b>61.88</b>  |
| <b>K-CAST</b>      |      |                |                 |                   |               |
| Llama 3.2          | 3b   | <b>92.60</b>   | <b>4.04</b>     | <b>22.92</b>      | <b>415.05</b> |
| 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 for both models.

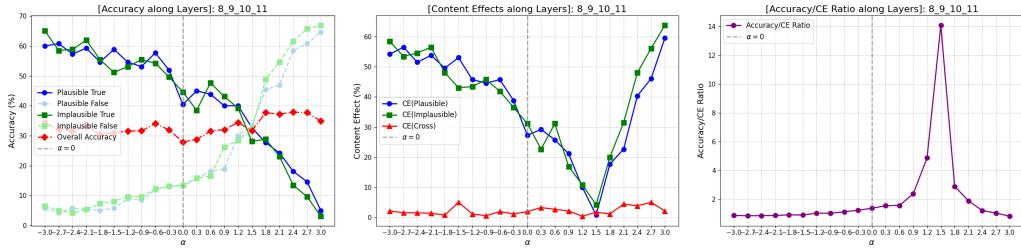


Figure 4: Robustness of steering to prompt variations on Llama 1b (i.e., ACC, CE, and ACC/CE). The results reveal that, despite some noise deriving from perturbations applied at test time, the overall effectiveness of steering remains unaltered.

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 3 (bottom) shows the impact of K-CAST on Llama 3b with different values of  $\alpha$  (with the sign dynamically determined at inference time).

### 4.3 Robustness to Prompt Perturbations

To test the robustness of steering performance to prompt perturbations [24], 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 [24], 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 [20, 10, 13]; (2) *Instruction induction*: inspired by [14], we provide five input-output pairs and ask GPT-4.5 to generate the possible instructions. Given a set of prompt variations (see Appendix), we compute the steering vectors using the original prompt and randomly select a variant at inference time.

**Steering is robust to prompt variations.** The results in Figure 4 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  $\alpha_{best}$  and some noise deriving from prompt variation, the overall effectiveness of steering remains unaltered. A similar trend is also observable for other models (see Appendix).

### 4.4 Impact of Steering 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

| Model     | Size | English                     |                                  |             | Chinese                     |                                  |             | German                      |                                  |             |
|-----------|------|-----------------------------|----------------------------------|-------------|-----------------------------|----------------------------------|-------------|-----------------------------|----------------------------------|-------------|
|           |      | $PPL_{\alpha=0} \downarrow$ | $PPL_{\alpha_{best}} \downarrow$ | $\Delta\%$  | $PPL_{\alpha=0} \downarrow$ | $PPL_{\alpha_{best}} \downarrow$ | $\Delta\%$  | $PPL_{\alpha=0} \downarrow$ | $PPL_{\alpha_{best}} \downarrow$ | $\Delta\%$  |
| Llama 3.2 | 1b   | 24.29                       | 24.77                            | <b>1.98</b> | 49.81                       | 51.35                            | <b>3.09</b> | 20.16                       | 20.52                            | <b>1.79</b> |
| 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   | <b>14.59</b>                | <b>15.17</b>                     | 3.98        | <b>18.41</b>                | <b>19.07</b>                     | 3.58        | <b>11.18</b>                | <b>11.58</b>                     | 3.57        |

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

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.

reasoning. For each experiment, we compare the performance between the model without steering ( $\alpha = 0$  in Table 3) against the model with steering ( $\alpha_{best}$ ). 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 [30] 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 [32] and Rulebreakers [6]) that were not presented during the steering modulating process. ProntoQA is a synthetic question-answering task designed to test deductive reasoning on multiple natural language premises. Rulebreakers is a question-answering 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 robust OOD generalization.

## 5 Conclusion

This paper investigated the problem of mitigating content biases on formal reasoning in LLMs through activation steering. Specifically, we curated a controlled syllogistic reasoning dataset to disentangle formal validity from content plausibility and systematically tested both static and conditional steering techniques on different families of LLMs.

Our results indicate that contrastive steering is particularly effective in reducing content effect and improving accuracy in formal reasoning, with a newly introduced kNN-based conditional steering method being able to achieve up to 15% absolute accuracy improvement.

In general, this paper demonstrates that activation-level interventions offer a scalable inference-time strategy for enhancing the robustness of LLMs, contributing towards more systematic and unbiased formal reasoning. In future work, we plan to explore how to enable better OOD generalization via steering for improving and debiasing more general reasoning capabilities in LLMs in different tasks and domains. To this end, we believe that conditional steering represents a promising solution via the dynamic adaptation of the steering modalities at test time. For this, different conditions could be explored to dynamically target different forms of reasoning.

## 6 Limitations

We acknowledge existing limitations with the current study:

The experiments are performed on LLMs of up to a size of 9 billion due to computational constraints. While the framework introduced in this paper is general and can in principle be applied to larger models, we left such an investigation to future work and to the broader research community.

The need to explicitly control and disentangle formal inference from plausibility led us to focus mainly on a synthetically generated task. While our study on language modeling and OOD generalization explores the impact of steering beyond our setup, we believe further work is still required to fully understand and improve the steering methodology, assessing how steering for content effect can impact real-world applications in critical domains (e.g., science, healthcare).

## References

- [1] Shikha Agarwal, Yifei Lyu, and Hongming Zhang. Mind the gap: From plausible to valid self-explanations in large language models. *arXiv preprint arXiv:2405.02706*, 2024.
- [2] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.
- [3] Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022.
- [4] Leonardo Bertolazzi, Albert Gatt, and Raffaella Bernardi. A systematic analysis of large language models as soft reasoners: The case of syllogistic inferences. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13882–13905, Miami, Florida, USA, November 2024. Association for Computational Linguistics.
- [5] Lorenzo Bertolazzi, Alberto Melloni, Marco Ferrari, and Mauro Dragoni. A systematic evaluation of logical reasoning in large language models. In *International Conference on Learning Representations (ICLR)*, 2024.
- [6] Jason Chan, Robert Gaizauskas, and Zhixue Zhao. Rulebreakers challenge: Revealing a blind spot in large language models’ reasoning with formal logic. *arXiv preprint arXiv:2410.16502*, 2024.
- [7] Ishita Dasgupta, Andrew Lampinen, Stephanie CY Chan, Antonia Creswell, James L McClelland, and Felix Hill. Content effects on logical reasoning in transformers. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.
- [8] Tiwalayo Eisape, Michael Tessler, Ishita Dasgupta, Fei Sha, Sjoerd Steenkiste, and Tal Linzen. 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)*, pages 8418–8437, 2024.
- [9] Deborah Ferreira, Julia Rozanova, Mokanarangan Thayaparan, Marco Valentino, and André Freitas. 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*, pages 194–201, 2021.
- [10] Hila Gonen, Srini Iyer, Terra Blevins, Noah A Smith, and Luke Zettlemoyer. Demystifying prompts in language models via perplexity estimation. *arXiv preprint arXiv:2212.04037*, 2022.
- [11] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.
- [12] Evan Hernandez, Belinda Z Li, and Jacob Andreas. Inspecting and editing knowledge representations in language models. In *First Conference on Language Modeling*.

- [13] Or Honovich, Thomas Scialom, Omer Levy, and Timo Schick. Unnatural instructions: Tuning language models with (almost) no human labor. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14409–14428, Toronto, Canada, July 2023. Association for Computational Linguistics.
- [14] Or Honovich, Uri Shaham, Samuel R. Bowman, and Omer Levy. Instruction induction: From few examples to natural language task descriptions. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1935–1952, Toronto, Canada, July 2023. Association for Computational Linguistics.
- [15] Geonhee Kim, Marco Valentino, and André Freitas. A mechanistic interpretation of syllogistic reasoning in auto-regressive language models. *arXiv preprint arXiv:2408.08590*, 2024.
- [16] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *arXiv preprint arXiv:2205.11916*, 2022.
- [17] Andrew K Lampinen, Ishita Dasgupta, Stephanie C Y Chan, Hannah R Sheahan, Antonia Creswell, Dharshan Kumaran, James L McClelland, and Felix Hill. Language models, like humans, show content effects on reasoning tasks. *PNAS Nexus*, 3(7):pgae233, 07 2024.
- [18] Guillaume Lample and François Charton. Deep learning for symbolic mathematics. *arXiv preprint arXiv:1912.01412*, 2019.
- [19] Bruce W. Lee, Inkit Padhi, Karthikeyan Natesan Ramamurthy, Erik Miehling, Pierre Dognin, Manish Nagireddy, and Amit Dhurandhar. Programming refusal with conditional activation steering. In *The Thirteenth International Conference on Learning Representations*, 2025.
- [20] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*, 2021.
- [21] Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems*, 36:41451–41530, 2023.
- [22] Qing Lyu, Shreya Havaldar, Adam Stein, Li Zhang, Delip Rao, Eric Wong, Marianna Apidianaki, and Chris Callison-Burch. Faithful chain-of-thought reasoning. In Jong C. Park, Yuki Arase, Baotian Hu, Wei Lu, Derry Wijaya, Ayu Purwarianti, and Adila Alfa Krisnadhi, editors, *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)*, pages 305–329, Nusa Dua, Bali, November 2023. Association for Computational Linguistics.
- [23] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.
- [24] Moran Mizrahi, Guy Kaplan, Dan Malkin, Rotem Dror, Dafna Shahaf, and Gabriel Stanovsky. State of what art? a call for multi-prompt LLM evaluation. *Transactions of the Association for Computational Linguistics*, 12:933–949, 2024.
- [25] Philipp Mondorf and Barbara Plank. 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)*, pages 9370–9402, 2024.
- [26] Liangming Pan, Alon Albalak, Xinyi Wang, and William Wang. Logic-LM: Empowering large language models with symbolic solvers for faithful logical reasoning. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3806–3824, Singapore, December 2023. Association for Computational Linguistics.
- [27] Nina Panickssery, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. Steering llama 2 via contrastive activation addition. *arXiv preprint arXiv:2312.06681*, 2023.

- [28] Xin Quan, Marco Valentino, Danilo S Carvalho, Dhairyा Dalal, and André Freitas. Peirce: Unifying material and formal reasoning via llm-driven neuro-symbolic refinement. *arXiv preprint arXiv:2504.04110*, 2025.
- [29] Xin Quan, Marco Valentino, Louise A. Dennis, and Andre Freitas. Verification and refinement of natural language explanations through LLM-symbolic theorem proving. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2933–2958, Miami, Florida, USA, November 2024. Association for Computational Linguistics.
- [30] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.
- [31] Leonardo Ranaldi, Marco Valentino, Alexander Polonsky, and André Freitas. Improving chain-of-thought reasoning via quasi-symbolic abstractions. *arXiv preprint arXiv:2502.12616*, 2025.
- [32] Abulhair Saparov and He He. Language models are greedy reasoners: A systematic formal analysis of chain-of-thought. In *The Eleventh International Conference on Learning Representations*, 2023.
- [33] John Seals, Saumik Dasgupta, Aishwarya Kumar, and Sandeep Ghosh. Do llms exhibit content effects? an investigation of human-like biases in language models. *CEUR Workshop Proceedings*, 3606:111–125, 2023.
- [34] S Seals and Valerie Shalin. 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)*, pages 8606–8622, 2024.
- [35] Nishant Subramani, Nivedita Suresh, and Matthew E Peters. Extracting latent steering vectors from pretrained language models. *arXiv preprint arXiv:2205.05124*, 2022.
- [36] Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussonot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.
- [37] Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*, 2023.
- [38] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.
- [39] Yuhuai Wu, Albert Qiaochu Jiang, Wenda Li, Markus Rabe, Charles Staats, Mateja Jamnik, and Christian Szegedy. Autoformalization with large language models. *Advances in Neural Information Processing Systems*, 35:32353–32368, 2022.
- [40] Magdalena Wysocka, Danilo Carvalho, Oskar Wysocki, Marco Valentino, and Andre Freitas. SyloBio-NLI: Evaluating large language models on biomedical syllogistic reasoning. In Luis Chiruzzo, Alan Ritter, and Lu Wang, editors, *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)*, pages 7235–7258, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics.
- [41] Yu Zhao, Alessio Devoto, Giwon Hong, Xiaotang Du, Aryo Pradipta Gema, Hongru Wang, Xuanli He, Kam-Fai Wong, and Pasquale Minervini. Steering knowledge selection behaviours in llms via sae-based representation engineering. *arXiv preprint arXiv:2410.15999*, 2024.
- [42] Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*, 2023.

## A Experimental Setup

We set the seed to 0 to enable a deterministic behaviour for the LLMs. We run all the steering experiments on a single A100 GPU. For probing, the graphs report the variance obtained with different random seeds (across 10 runs).

## B Linear Probing Results

Results are reported in Figure 5.

## C Conditional Steering Results

A detailed comparison between CAST and K-CAST is reported in Figure 6 and Figure 7.

## D Syllogistic Schemes

Here is the list of syllogistic schemes adopted to construct the dataset. The schemes are aligned with previous work on syllogistic reasoning with LLMs [4].

Schema: AA1

Premise 1: All  $\langle A \rangle$  are  $\langle B \rangle$  / Premise 2: All  $\langle B \rangle$  are  $\langle C \rangle$

Conclusions: All  $\langle A \rangle$  are  $\langle C \rangle$  | some  $\langle A \rangle$  are  $\langle C \rangle$  | some  $\langle C \rangle$  are  $\langle A \rangle$

Schema: AA2

Premise 1: All  $\langle B \rangle$  are  $\langle A \rangle$  / Premise 2: All  $\langle C \rangle$  are  $\langle B \rangle$

Conclusions: All  $\langle C \rangle$  are  $\langle A \rangle$  | some  $\langle A \rangle$  are  $\langle C \rangle$  | some  $\langle C \rangle$  are  $\langle A \rangle$

Schema: AA4

Premise 1: All  $\langle B \rangle$  are  $\langle A \rangle$  / Premise 2: All  $\langle B \rangle$  are  $\langle C \rangle$

Conclusions: Some  $\langle A \rangle$  are  $\langle C \rangle$  | some  $\langle C \rangle$  are  $\langle A \rangle$

Schema: AI2

Premise 1: All  $\langle B \rangle$  are  $\langle A \rangle$  / Premise 2: Some  $\langle C \rangle$  are  $\langle B \rangle$

Conclusions: Some  $\langle A \rangle$  are  $\langle C \rangle$  | some  $\langle C \rangle$  are  $\langle A \rangle$

Schema: AI4

Premise 1: All  $\langle B \rangle$  are  $\langle A \rangle$  / Premise 2: Some  $\langle B \rangle$  are  $\langle C \rangle$

Conclusions: Some  $\langle A \rangle$  are  $\langle C \rangle$  | some  $\langle C \rangle$  are  $\langle A \rangle$

Schema: AO3

Premise 1: All  $\langle A \rangle$  are  $\langle B \rangle$  / Premise 2: Some  $\langle C \rangle$  are not  $\langle B \rangle$

Conclusions: Some  $\langle C \rangle$  are not  $\langle A \rangle$

Schema: AO4

Premise 1: All  $\langle B \rangle$  are  $\langle A \rangle$  / Premise 2: Some  $\langle B \rangle$  are not  $\langle C \rangle$

Conclusions: Some  $\langle A \rangle$  are not  $\langle C \rangle$

Schema: AE1

Premise 1: All  $\langle A \rangle$  are  $\langle B \rangle$  / Premise 2: No  $\langle B \rangle$  are  $\langle C \rangle$

Conclusions: No  $\langle A \rangle$  are  $\langle C \rangle$  | no  $\langle C \rangle$  are  $\langle A \rangle$  |  
some  $\langle A \rangle$  are not  $\langle C \rangle$  | some  $\langle C \rangle$  are not  $\langle A \rangle$

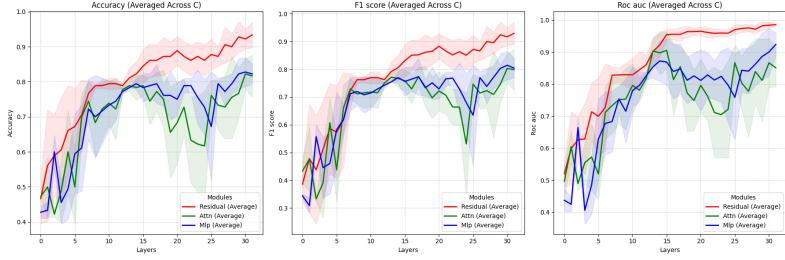
Schema: AE2

Premise 1: All  $\langle B \rangle$  are  $\langle A \rangle$  / Premise 2: No  $\langle C \rangle$  are  $\langle B \rangle$

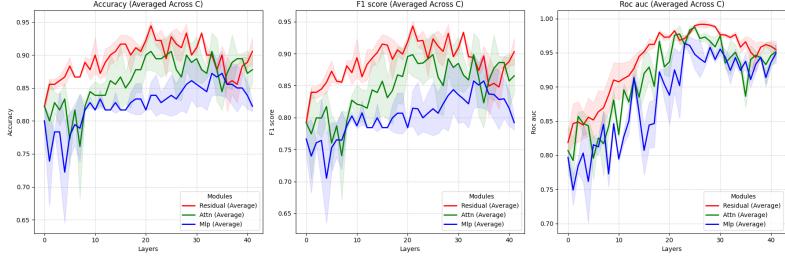
Conclusions: Some  $\langle A \rangle$  are not  $\langle C \rangle$

Schema: AE3

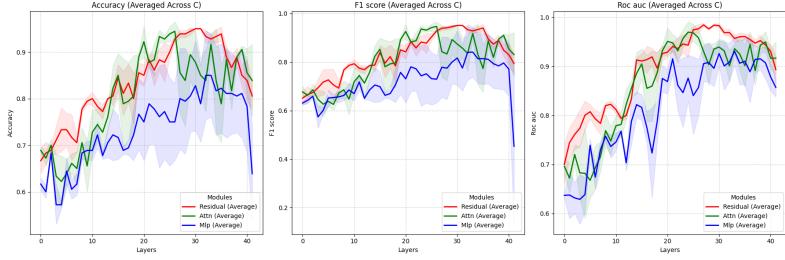
Premise 1: All  $\langle A \rangle$  are  $\langle B \rangle$  / Premise 2: No  $\langle C \rangle$  are  $\langle B \rangle$



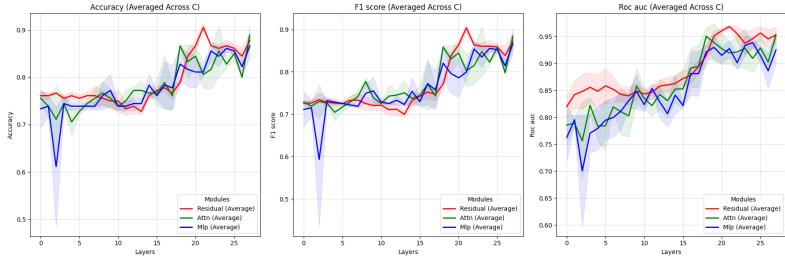
(a) Plausibility – Llama



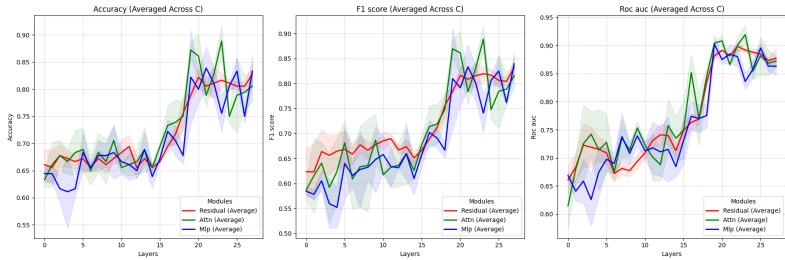
(b) Validity – Gemma 9b



(c) Plausibility – Gemma 9b



(d) Validity – Qwen 7b



(e) Validity – Qwen 7b

Figure 5: Linear Probing Results for different models. In general, the probing results suggest that the information about validity and plausibility is encoded in the second half of the layers, with a peak in the third quarter (predominantly in the residual stream).

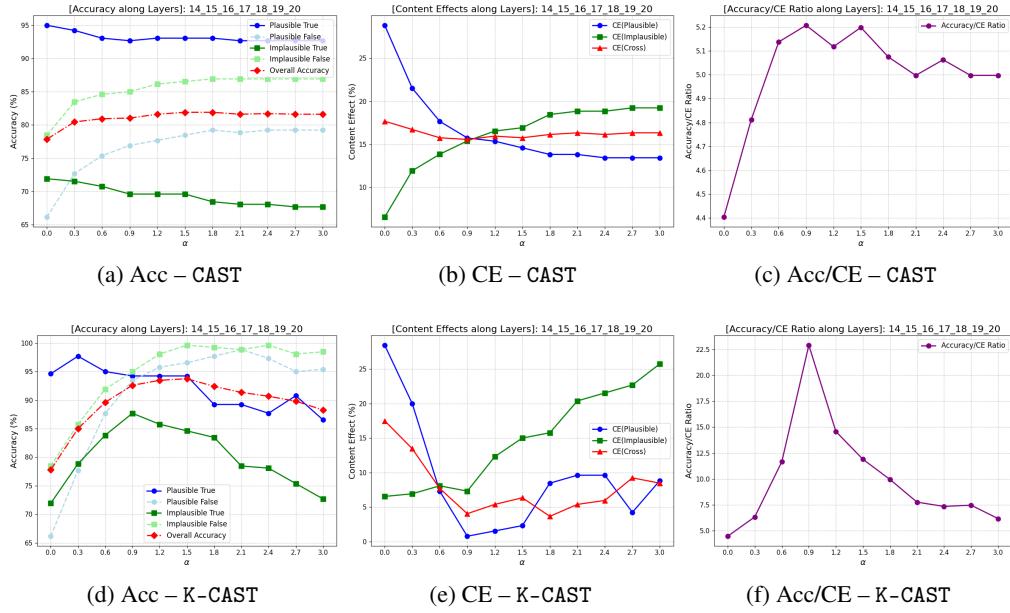


Figure 6: Results of conditional activation steering on Llama-3.2-3b-Instruct. Standard conditional steering (top) and KNN-based conditional steering (bottom).

Conclusions: No  $\langle A \rangle$  are  $\langle C \rangle$  | no  $\langle C \rangle$  are  $\langle A \rangle$  | some  $\langle A \rangle$  are not  $\langle C \rangle$  | some  $\langle C \rangle$  are not  $\langle A \rangle$

Schema: AE4

Premise 1: All  $\langle B \rangle$  are  $\langle A \rangle$  / Premise 2: No  $\langle B \rangle$  are  $\langle C \rangle$   
 Conclusions: Some  $\langle A \rangle$  are not  $\langle C \rangle$

Schema: IA1

Premise 1: Some  $\langle A \rangle$  are  $\langle B \rangle$  / Premise 2: All  $\langle B \rangle$  are  $\langle C \rangle$   
 Conclusions: Some  $\langle A \rangle$  are  $\langle C \rangle$  | some  $\langle C \rangle$  are  $\langle A \rangle$

Schema: IA4

Premise 1: Some  $\langle B \rangle$  are  $\langle A \rangle$  / Premise 2: All  $\langle B \rangle$  are  $\langle C \rangle$   
 Conclusions: Some  $\langle A \rangle$  are  $\langle C \rangle$  | some  $\langle C \rangle$  are  $\langle A \rangle$

Schema: IE1

Premise 1: Some  $\langle A \rangle$  are  $\langle B \rangle$  / Premise 2: No  $\langle B \rangle$  are  $\langle C \rangle$   
 Conclusions: Some  $\langle A \rangle$  are not  $\langle C \rangle$

Schema: IE2

Premise 1: Some  $\langle B \rangle$  are  $\langle A \rangle$  / Premise 2: No  $\langle C \rangle$  are  $\langle B \rangle$   
 Conclusions: Some  $\langle A \rangle$  are not  $\langle C \rangle$

Schema: IE3

Premise 1: Some  $\langle A \rangle$  are  $\langle B \rangle$  / Premise 2: No  $\langle C \rangle$  are  $\langle B \rangle$   
 Conclusions: Some  $\langle A \rangle$  are not  $\langle C \rangle$

Schema: IE4

Premise 1: Some  $\langle B \rangle$  are  $\langle A \rangle$  / Premise 2: No  $\langle B \rangle$  are  $\langle C \rangle$   
 Conclusions: Some  $\langle A \rangle$  are not  $\langle C \rangle$

Schema: OA3

Premise 1: Some  $\langle A \rangle$  are not  $\langle B \rangle$  / Premise 2: All  $\langle C \rangle$  are  $\langle B \rangle$

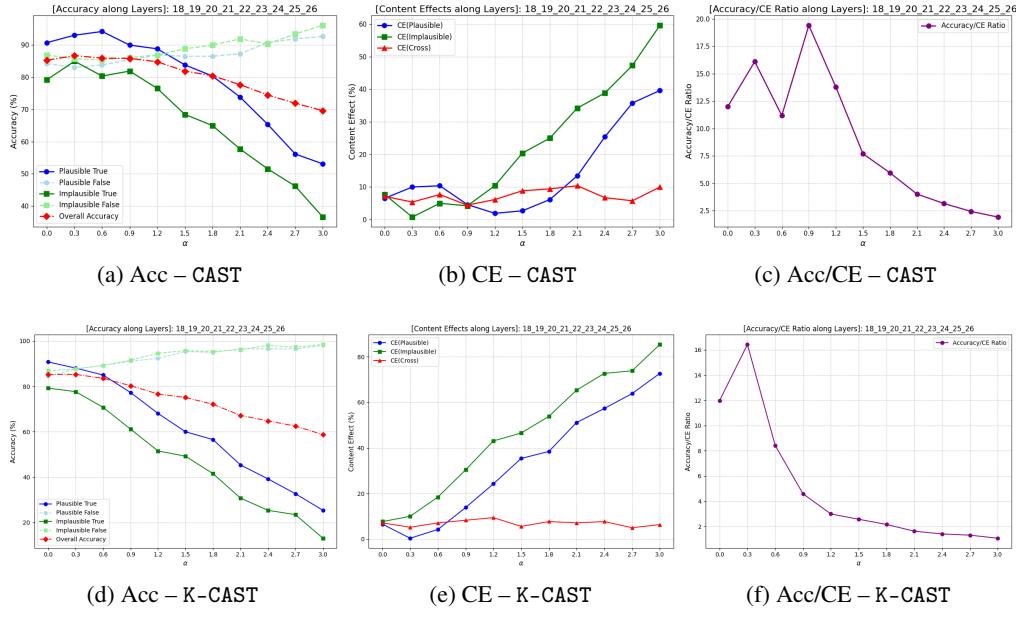


Figure 7: Results of conditional activation steering on Qwen 2.5 3b. Standard conditional steering (top) and KNN-based conditional steering (bottom).

Conclusions: Some  $\langle A \rangle$  are not  $\langle C \rangle$

Schema: OA4

Premise 1: Some  $\langle B \rangle$  are not  $\langle A \rangle$  / Premise 2: All  $\langle B \rangle$  are  $\langle C \rangle$   
 Conclusions: Some  $\langle C \rangle$  are not  $\langle A \rangle$

Schema: EA1

Premise 1: No  $\langle A \rangle$  are  $\langle B \rangle$  / Premise 2: All  $\langle B \rangle$  are  $\langle C \rangle$   
 Conclusions: Some  $\langle C \rangle$  are not  $\langle A \rangle$

Schema: EA2

Premise 1: No  $\langle B \rangle$  are  $\langle A \rangle$  / Premise 2: All  $\langle C \rangle$  are  $\langle B \rangle$   
 Conclusions: No  $\langle A \rangle$  are  $\langle C \rangle$  | no  $\langle C \rangle$  are  $\langle A \rangle$  |  
 some  $\langle A \rangle$  are not  $\langle C \rangle$  | some  $\langle C \rangle$  are not  $\langle A \rangle$

Schema: EA3

Premise 1: No  $\langle A \rangle$  are  $\langle B \rangle$  / Premise 2: All  $\langle C \rangle$  are  $\langle B \rangle$   
 Conclusions: No  $\langle A \rangle$  are  $\langle C \rangle$  | no  $\langle C \rangle$  are  $\langle A \rangle$  |  
 some  $\langle A \rangle$  are not  $\langle C \rangle$  | some  $\langle C \rangle$  are not  $\langle A \rangle$

Schema: EA4

Premise 1: No  $\langle B \rangle$  are  $\langle A \rangle$  / Premise 2: All  $\langle B \rangle$  are  $\langle C \rangle$   
 Conclusions: Some  $\langle C \rangle$  are not  $\langle A \rangle$

Schema: EI1

Premise 1: No  $\langle A \rangle$  are  $\langle B \rangle$  / Premise 2: Some  $\langle B \rangle$  are  $\langle C \rangle$   
 Conclusions: Some  $\langle C \rangle$  are not  $\langle A \rangle$

Schema: EI2

Premise 1: No  $\langle B \rangle$  are  $\langle A \rangle$  / Premise 2: Some  $\langle C \rangle$  are  $\langle B \rangle$   
 Conclusions: Some  $\langle C \rangle$  are not  $\langle A \rangle$

Schema: EI3

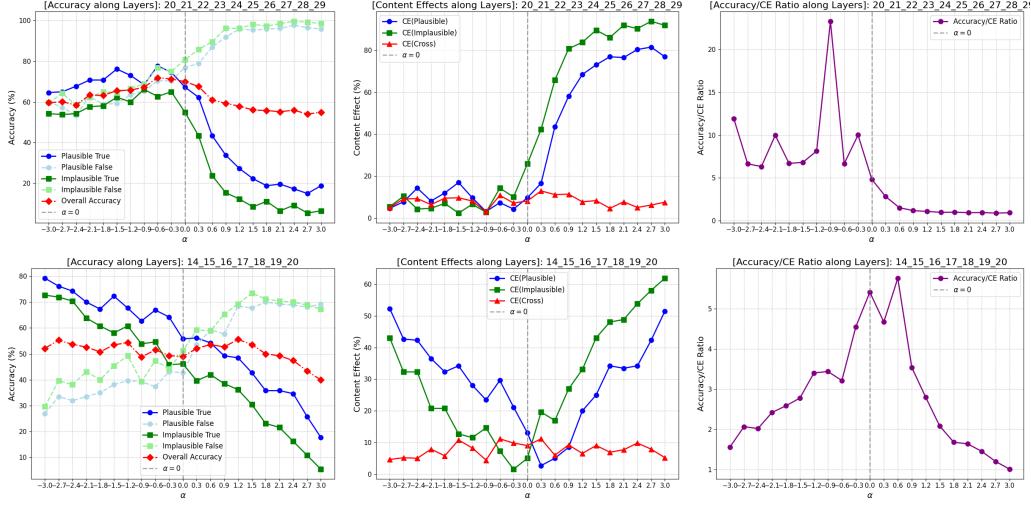


Figure 8: Robustness of steering to prompt perturbations on Gemma 2 9b (top) and Qwen 2.5 7b (bottom) (i.e., ACC, CE, and ACC/CE from left to right). Steering is still effective despite variations applied to the prompts at test time.

Premise 1: No <A> are <B> / Premise 2: Some <C> are <B>  
Conclusions: Some <C> are not <A>

Schema: EI4

Premise 1: No <B> are <A> / Premise 2: Some <B> are <C>  
Conclusions: Some <C> are not <A>

## E Robustness to Prompt Perturbations

Results are reported in Figure 8.

## F Prompts Templates

Variation 0:

Given the premises , evaluate the validity of the conclusion .

Premises :

```
entry [ ‘premise1’ ].  
entry [ ‘premise2’ ].
```

Conclusion: entry [ ‘conclusion’ ].

The conclusion is

Variation 1:

Carefully evaluate the validity of the following logical argument  
and answer ’the argument is logically valid’ or ’the argument is logically invalid’ .

Premise1: entry [ ‘premise1’ ].

Premise2: entry [ ‘premise2’ ].

Conclusion: entry [ ‘conclusion’ ].

The argument is logically

Variation 2:

Analyze the formal logical structure of the argument below, then indicate whether it is valid or invalid.

Premise 1: entry['premise1']\n

Premise 2: entry['premise2']\n

Conclusion: entry['conclusion']\n\n

The logical structure is

Variation 3:

Assess carefully whether the argument presented below is logically valid or invalid based on the given premises and conclusion.

1. entry['premise1']
2. entry['premise2']

Conclusion:

entry['conclusion']

The argument is logically

Variation 4:

Examine the following argument, generating "valid" if the conclusion logically follows from the premises provided, "invalid" otherwise.

- Premise 1: entry['premise1']

- Premise 2: entry['premise2']

Conclusion: entry['conclusion']

The argument is

Variation 5:

Given the two premises and the conclusion below, judge carefully whether the logical argument is valid or invalid.

Premise 1: entry['premise1']

Premise 2: entry['premise2']

Conclusion: entry['conclusion']

The logical argument is

Variation 6:

Evaluate the following logical argument carefully and decide whether the provided conclusion is formally valid or invalid given the two premises.

Premise 1: entry['premise1']

Premise 2: entry['premise2']

Conclusion: entry['conclusion']

The conclusion is formally