How Language Models Conflate Logical Validity with Plausibility: A Representational Analysis of Content Effects

Anonymous ACL submission

Abstract

Both humans and large language models (LLMs) exhibit *content effects*: biases in which the plausibility of the semantic content of a reasoning problem influences judgments regarding its logical validity. While this phenomenon in humans is best explained by the dual-process theory of reasoning, the mechanisms behind content effects in LLMs remain unclear. In this work, we address this issue by investigating how LLMs encode the concepts of validity and plausibility within their internal representations. We show that both concepts are linearly represented and strongly aligned in representational geometry, leading models to conflate plausibility with validity. Using steering vectors, we demonstrate that plausibility vectors can causally bias validity judgements, and vice versa, and that the degree of alignment between these two concepts predicts the magnitude of behavioral content effects across models. Finally, we construct debiasing vectors that disentangle these concepts, reducing content effects and improving reasoning accuracy. Our findings advance understanding of how abstract logical concepts are represented in LLMs and highlight representational interventions as a path toward more logical systems.

1 Introduction

A pure abstract reasoner applies logical rules and manipulates symbols independently of the content or context in which they appear. Both humans and LLMs often deviate from this ideal, exhibiting systematic biases where content influences formal reasoning.

Content effects are well-documented in human reasoning tasks, such as judging the validity of syllogistic inferences, where syllogisms with plausible conclusions are often judged as valid despite being logically incorrect (Evans et al., 1983). Among the theories proposed to account for such biases, the dual-process theory of reasoning has been the most influential, positing two distinct modes of thought: a fast, intuitive, heuristic-driven system (System 1), and a slower, deliberative system responsible for analytical reasoning (System 2; Evans, 2008; Kahneman, 2011). Neuroscientific studies have provided empirical support for this framework, highlighting different neural substrates associated with these reasoning processes (Goel et al., 2000; Stollstorff et al., 2012). Recent work has found that LLMs exhibit similar content effects in reasoning tasks (Lampinen et al., 2024); however, the underlying mechanisms driving these effects remain unknown.

In this work, we provide a representational account of why content effects may emerge in LLMs, investigating how the abstract concepts of validity and plausibility are encoded in their hidden representation space. Specifically, we build upon the linear representation hypothesis (Park et al., 2024), which proposes that many high-level concepts are encoded linearly within the latent space of LLMs. This hypothesis has been supported by empirical findings showing that concepts can often be captured by linear probes or manipulated with steering vectors (Liu et al., 2024; Rimsky et al., 2024; Marks and Tegmark, 2024). We hypothesize that content effects in LLMs may arise from the way validity and plausibility are entangled within the model's representational geometry. Specifically, we predict that LLMs conflate validity with plausibility, leading to systematic biases in reasoning. To test this hypothesis, we analyze ten different LLMs using zero-shot and chain-of-thought (CoT) prompting (Wei et al., 2022) and address the following research questions:

RQ1: *Do current LLMs exhibit content effects?* We find that models from the Qwen-2.5 (Yang et al., 2025b), Qwen-3 (Yang et al., 2025a), and Gemma-3 (Kamath et al., 2025) families display systematic biases where plausibility influences validity judgments.

RQ2: *How do LLMs encode plausibility and*

	Plausible	Implausible
Valid	**P1:** All labradors are canines. **P2:** All labradors are dogs. **C:** *Some dogs are canines.*	**P1:** All canines are cats. **P2:** All canines are dogs. **C:** *Some dogs are cats.*
Invalid	**P1:** No dogs are cats. **P2:** No canines are cats. **C:** *Some dogs are canines.*	**P1:** All dogs are labradors. **P2:** Some labradors are not canines. **C:** *No dogs are canines.*

Figure 1: **Validity and plausibility configurations.** Illustrative examples of valid and invalid syllogisms with plausible and implausible conclusions. Here, *plausible* indicates that the conclusion is true in the real world, whereas *implausible* indicates that it is false.

validity? We find single vectors that can control models' judgements for both validity and plausibility. Moreover, these vectors are highly similar.

RQ3: *What do the representations reveal about behavioral content effects?* We demonstrate that greater geometric similarity between validity and plausibility vectors is correlated with stronger behavioral content effects across models. Moreover, we establish causal interaction across concepts: plausibility vectors steers validity judgments, and vice versa.

RQ4: *Can we design an intervention to mitigate content effects?* We develop debiasing steering vectors that disentangle validity from plausibility, reducing content effects while improving reasoning accuracy.[1]

2 Related Work

2.1 Content Effects in Humans and LLMs

Content effects describe the well-documented tendency in humans to evaluate reasoning problems based on their semantic content and prior beliefs rather than logical structure (Markovits and Nantel, 1989). This phenomenon has been extensively studied in syllogistic reasoning (Evans et al., 1983; Oakhill and Johnson-Laird, 1985), the Wason selection task (Wason, 1968), and Bayesian inference, where existing beliefs can override statistical evidence (Kahneman and Tversky, 1973; Bar-Hillel, 1980).

Recent research has revealed that LLMs exhibit similar content effects. Lampinen et al. (2024) found that LLMs exhibit this bias in the Wason selection task and syllogistic inference. Bertolazzi et al. (2024) showed that in a multiple-choice setting with plausible and implausible conclusions, LLMs favor the former as valid conclusions in syllogisms, regardless of their logical validity. Balappanawar et al. (2025) demonstrated that content effects in LLMs extend beyond conditional reasoning and syllogistic inference to various inference rules in propositional logic, suggesting a broader pattern of belief-based reasoning. While this research provides robust behavioral evidence for content effects in LLMs, the underlying mechanisms governing this bias remain poorly understood.

2.2 Linear Representations in LLMs

Recent work suggests that many high-level concepts are encoded linearly in the latent space of LLMs and can be manipulated using steering vectors (Liu et al., 2024). For instance, Zhao et al. (2025) identified a harmfulness direction, where steering along this dimension causes LLMs to interpret harmless instructions as harmful. Similarly, Marks and Tegmark (2024) show that truth is linearly represented in LLMs, with interventions along the truthful direction leading models to treat false statements as true, and vice versa. Other concepts, including sentiment (Tigges et al., 2024) and refusal (Arditi et al., 2025), have also been found to be linearly encoded within LLMs, among others. At the same time, recent work has also shown that not all concepts adhere to the linear representation hypothesis, with some high-level features requiring more complex, non-linear encoding structures (Engels et al., 2025).

While the concepts that interest us the most—truth (Marks and Tegmark, 2024) and log-

[1] Code and data will be made available, CC BY license, upon acceptance.

ical validity (Valentino et al., 2025)—have been shown to be linearly represented in the latent space of LLMs, no study has yet examined how content effects in LLMs might emerge from the interaction between these two dimensions.

3 Method

3.1 Datasets and Tasks

We focus on syllogistic inferences, a type of logical problem in which LLMs exhibit content effects (Lampinen et al., 2024; Bertolazzi et al., 2024).[2]

Data. Starting from the data from Bertolazzi et al. (2024), we construct 1,280 syllogisms, each containing two premises and a conclusion. We utilize this dataset because it systematically covers all 64 types of syllogism across two distinct conditions: "plausible" and "implausible" ones. The plausible syllogisms have conclusions that are true in the actual world (e.g., "Some dogs are canines"), while implausible syllogisms' conclusions are false in the actual world (e.g., "Some dogs are cats"). The dataset is constructed using ten distinct triples of terms exhibiting a taxonomical relationship (e.g., "labradors" → "dogs" → "canines"). We provide further details on the data generation process in Appendix B.

Logical validity classification task. Each syllogism can be classified as either valid or invalid and has a conclusion that can be either plausible or implausible. This creates four distinct categories of syllogisms that enable us to examine the interaction between logical structure and content plausibility. Figure 1 provides illustrative examples of each combination of validity and plausibility labels. Across the experiments we perform, we partition this dataset using a 70-30 train-test split. Models are tasked with performing syllogistic reasoning as a binary classification problem, where they must determine the logical validity of presented syllogisms.

Plausibility classification task. Plausibility can be interpreted as an inherently continuous and nuanced concept. However, to enable comparison with logical validity, we operationalize it as a binary notion. Throughout this paper, we define plausibility operationally as the truth value of a statement with respect to the actual world: a statement is considered plausible if it is factually true. We extract all unique conclusions present in the original data from Bertolazzi et al. (2024) and construct a task requiring models to classify these statements as either true or false based on what models "believe" to be factually accurate.[3]

Auxiliary tasks. We additionally incorporate control datasets for two auxiliary binary classification tasks. The first task requires, given a source and a target term in a taxonomical relationship, to determine whether the source is a hypernym or a hyponym of the target. Crucially, this dataset is built using the same ten triples that were used in the two binary classification tasks above. The second task involves harmful/harmless content classification using the data from Arditi et al. (2025).[4] These auxiliary binary classification tasks serve as experimental controls to verify that any observed interactions between plausibility and validity are not merely artifacts of both tasks being binary classifications or of sharing similar vocabulary and prompt structures.

3.2 Representing Binary Concepts as Single Directions

We represent each binary concept as a linear direction in the model's activation space using the difference-in-means approach (Belrose, 2023; Rimsky et al., 2024; Marks and Tegmark, 2024). For a binary concept with a positive class (e.g., valid, plausible, harmless) and a negative class (e.g., invalid, implausible, harmful), we compute the mean activation vectors for each class at layer l and at the *last token* position right before the label prediction. Importantly, we use the model's predicted labels rather than ground-truth labels to define class membership, as our focus is on understanding the model's own internal "beliefs". The concept direction is then defined as:

$$v^l_{\text{concept}} = \mu^l_{\text{positive}} - \mu^l_{\text{negative}}$$

where μ^l_{positive} and μ^l_{negative} are the mean hidden activations of the positive and negative classes, respectively. This formulation defines a direction in activation space that points from the negative class toward the positive class.

[2]For a detailed explanation of the structure of syllogisms, see Appendix A.

[3]The complete prompt formulations for the logical validity and the plausibility tasks are provided in Appendix B.

[4]This dataset consists of a set of prompts or instructions, and an LLM is tasked with judging whether they are harmful or harmless. See Appendix B for additional details on the data of these two control datasets.

3.3 Steering Approach

Once we extract a vector representing a binary concept, we can then use it as a steering vector (Liu et al., 2024). We manipulate the model's behavior by adding or subtracting the vector from the same layer l from which it was extracted: adding v^l_{concept} to the activations steers the model toward the positive class (e.g., making outputs more valid, plausible, or harmless), while subtracting it steers toward the negative class (e.g., making outputs more invalid, implausible, or harmful). This bidirectional control allows us to test whether the concept is linearly represented at each layer by measuring how effectively these interventions change the model's classification behavior. Across our steering experiments, each vector is added or subtracted as is, without multiplying it by a scalar or normalizing, which means that we expect the vector to represent the concept faithfully in both direction and magnitude.

3.4 Metrics

We use the following metrics in our experiments.

Content effect. When evaluating models on the logical validity classification task, we need a way to precisely quantify the degree of content effect. To this end, we introduce a metric that enables fine-grained comparisons across models. Let $A(S)$ denote the accuracy of a model on a subset of the logical validity dataset $S \subseteq D$. We partition the dataset D by validity (v^+ for valid, v^- for invalid) and plausibility (p^+ for plausible, p^- for implausible), yielding four disjoint subsets $D_{v^+,p^+}, D_{v^+,p^-}, D_{v^-,p^+}, D_{v^-,p^-}$. We then define:

$$\begin{aligned}\Delta_{v^+} &= A(D_{v^+,p^+}) - A(D_{v^+,p^-}),\\ \Delta_{v^-} &= A(D_{v^-,p^-}) - A(D_{v^-,p^+})\end{aligned}$$

Here, Δ_{v^+} captures how much more accurately the model classifies valid arguments when conclusions are plausible, while Δ_{v^-} captures the corresponding effect for invalid arguments with implausible conclusions.

The overall *content effect* (CE) is the mean of these two components:

$$\text{CE} = \frac{1}{2}(\Delta_{v^+} + \Delta_{v^-}).$$

This metric is bounded between -1 and 1: $\text{CE} = 1$ indicates judgments of validity are fully driven by plausibility, $\text{CE} = 0$ indicates independence of validity and plausibility.[5]

Steering power. Since we work with binary concepts and apply steering vectors to control model behavior on binary classification tasks, we introduce *steering power* as a measure of the effectiveness of such vectors at a given layer l. Following the approach described above, we always steer against the model's original prediction: *adding* v^l when the model predicts negative, and *subtracting* it when the model predicts positive.

Formally, the *steering power* (SP) of a vector v^l is the proportion of examples whose predicted label flips when the signed vector is applied at layer l:

$$\text{SP}(v; \mathcal{D}) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\Big[\hat{y}_i' \neq \hat{y}_i\Big]$$

where $\hat{y}_i$ is the original prediction for input x_i, and $\hat{y}_i'$ is the steered prediction.

3.5 Models

We ran experiments using instruction-tuned versions of three variants of LLMs: Qwen-2.5 (Yang et al., 2025b), Qwen-3 (Yang et al., 2025a), and Gemma-3 (Kamath et al., 2025), with parameter counts ranging from 4B to 32B. This choice combines both standard instruction-tuned models (Qwen-2.5 and Gemma-3) and newer "thinking" models (Qwen-3) that have been trained to produce long reasoning traces during inference.

Across all experiments, we evaluate models using both zero-shot prompting and CoT prompting. We primarily present results using `Qwen2.5-32B-Instruct` and `Qwen3-14B` as our main reference models. We observe consistent patterns across all evaluated models, with full comparative results in Appendix D.

4 Experiments and Results

4.1 Do LLMs Exhibit Content Effects?

We ask whether LLMs suffer from content effects on the logical validity classification task. We expect biased LLMs to classify more accurately syllogisms where plausibility supports logical validity: that is, when the conclusion is plausible and the argument is valid, or when the conclusion is implausible and the argument is invalid. Conversely,

[5] A negative CE would instead indicate bias in the opposite direction.

	Qwen2.5-32B		Qwen3-14B	
	0-shot	**CoT**	**0-shot**	**CoT**
D_{v^+,p^+}	100.00	98.67	97.33	95.31
D_{v^-,p^+}	67.50	86.64	90.83	99.10
D_{v^+,p^-}	60.92	93.10	60.92	92.50
D_{v^-,p^-}	98.04	100.00	97.06	99.90
Acc	81.62	94.60	86.54	96.70
CE	0.348	0.096	0.213	0.014

Table 1: **Behavioral performance.** Zero-shot vs CoT accuracy for `Qwen2.5-32B-Instruct` and `Qwen3-14B` on different subsets of syllogism dataset and overall content effect. Subsets are organized by validity label (v^+ = valid, v^- = invalid) and plausibility of the conclusion (p^+ = plausible, p^- = implausible).

models should have lower accuracy judging the validity of other cases.

Table 1 reports the accuracies of `Qwen2.5-32B-Instruct` and `Qwen3-14B` on the test split of the logical validity classification task. Both models exhibit content effects in the zero-shot setting, while the bias is greatly reduced with CoT prompting. Looking at the CE metric, we observe that for `Qwen2.5-32B-Instruct` it is relatively high in zero-shot (0.348), showing strong conflation of plausibility with validity. When prompted with CoT, CE drops substantially to 0.096, indicating that explicit reasoning almost eliminates the bias. For `Qwen3-14B`, the zero-shot CE is lower (0.213), suggesting less susceptibility to content effects even without structured reasoning. When given CoT prompting, CE drops almost to zero (0.014), implying that validity judgments become nearly independent of plausibility.

Across all models, CoT prompting achieves a significantly lower CE compared to zero-shot (Mann-Whitney U test, $p < 0.01$; see the complete behavioral results in Table 4 in the Appendix).

4.2 How are Plausibility and Validity Encoded in LLMs?

We now turn to how validity and plausibility are represented internally by LLMs.

Single directions control validity and plausibility judgements. To test the linear representation hypothesis on validity and plausibility, we conduct steering experiments to identify layers where these concepts are encoded linearly. For each hidden layer l, we compute the difference-in-means vectors

$$v^l_{\text{validity}} = \mu^l_{v^+} - \mu^l_{v^-}, \quad v^l_{\text{plausibility}} = \mu^l_{p^+} - \mu^l_{p^-}$$

representing validity and plausibility directions at layer l. Analogously, using the auxiliary datasets (harmful/harmless, and hypernym/hyponym), we obtain $v^l_{\text{harmlessness}} = \mu^l_{harm^+} - \mu^l_{harm^-}$ and $v^l_{\text{hypernymy}} = \mu^l_{hyp^+} - \mu^l_{hyp^-}$, respectively.

Figure 2a shows SP values across all 63 layers of `Qwen2.5-32B-Instruct` in both zero-shot and CoT prompting. Both validity and plausibility vectors achieve SP ≈ 1 at late layers, but near zero at early layers. Steering is equally effective in zero-shot and CoT, and both tasks peak at similar layers. This demonstrates that logical validity and plausibility classification can be effectively controlled by single directions in the model's latent space.

Validity and plausibility vectors are similar. Having identified the late layers where steering is effective, we now examine whether validity and plausibility are represented similarly. Figure 2b provides a qualitative visualization through a 3D PCA projection of hidden states from layer 50 (of 64) of `Qwen2.5-32B-Instruct` using data from both validity and plausibility classification in the zero-shot setting. In the logical validity task, the model classifies arguments as valid or invalid, while in the plausibility task, it classifies statements as true or false.[6] At this late layer, four well-separated clusters emerge. Notably, the plausibility clusters (true/false) are displaced along a similar direction as validity clusters (valid/invalid), indicating parallel representational structure.[7]

To quantitatively verify the similarity observed in the visualization, we compute the average cosine similarity between validity and plausibility steering vectors. Specifically, we restrict to layers with SP > 0.75 (highlighted in yellow in Figure 2a), since steering vectors are meaningful only when they can effectively control predictions. Because both validity and plausibility concepts are extracted from prompts that instantiate binary classification tasks and have lexical overlap, we also compute similarities between validity and harmlessness, and between validity and hypernymy, as

[6] Labels in the plot are model predictions rather than ground truth, since we are interested in the model's internal representation of what it "believes" to be valid/invalid or true/false.

[7] We include visualizations across layers for both the zero-shot and CoT settings in the Appendix D.2.

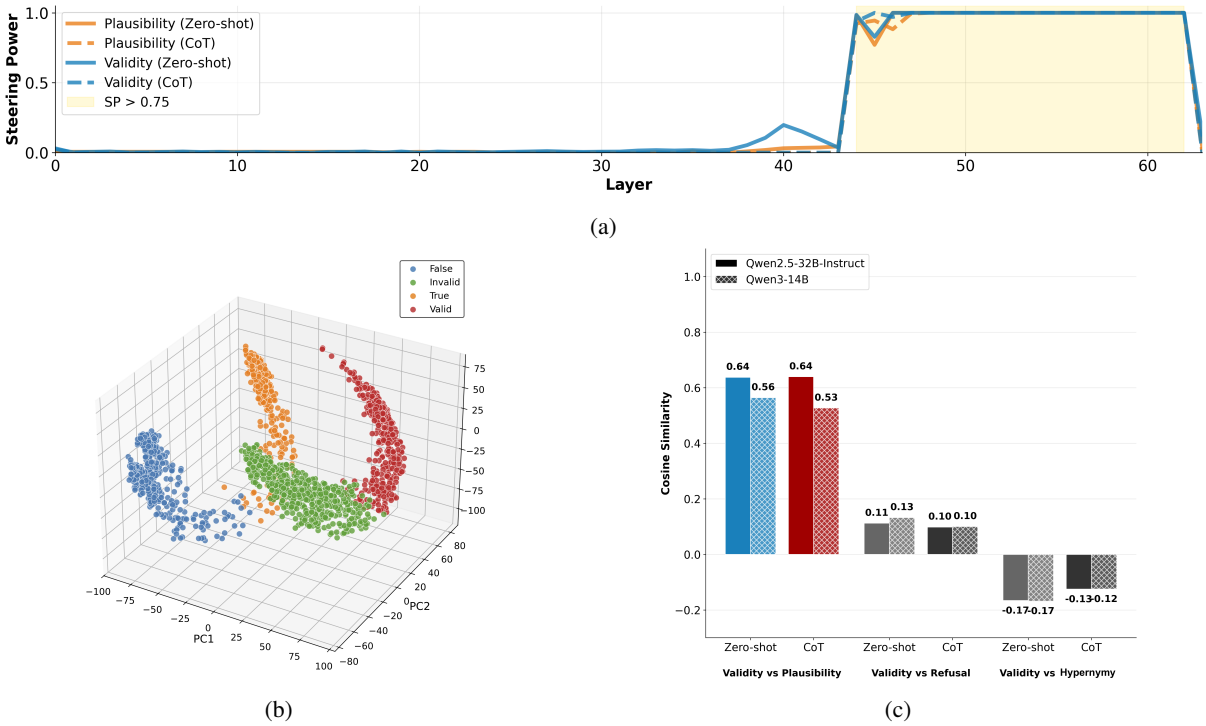


Figure 2: **Representational analysis of validity and plausibility concepts**. (a) Steering power (SP) of validity and plausibility vectors applied at different hidden layers of `Qwen2.5-32B-Instruct`. The region in yellow highlights layers with SP > 0.75. Validity and plausibility steering vectors show high SP at similar layers using both zero-shot and CoT prompting. (b) 3D PCA projection of hidden states from layer 50 of `Qwen2.5-32B-Instruct` in the zero-shot setting showing four distinct clusters corresponding to model predictions (valid/invalid, true/false). The parallel geometric structure between true/false and valid/invalid clusters suggests shared representational directions for plausibility and validity. (c) Average cosine similarity between the validity vector and vectors for the concepts of plausibility, hypernymy, and harmlessness across all layers for both models under zero-shot and CoT prompting. High validity-plausibility alignment (0.53 to 0.64) contrasts with low alignment for other concepts (0.10 to 0.13 and -0.12 to -0.17), confirming specific representational entanglement.

controls. Figure 2c reports the average cosine similarities across layers for `Qwen2.5-32B-Instruct` and `Qwen3-14B` under zero-shot and CoT prompting. For both models, validity and plausibility vectors are substantially aligned (cosine similarity 0.48 to 0.64), while the two other concepts have a significantly lower similarity (0.10 to 0.13 and -0.12 to -0.17). This indicates that the observed similarity is specific to validity and plausibility. Across all evaluated models, unlike CE, the degree of alignment between validity and plausibility vectors does not differ significantly between prompting conditions (Mann-Whitney U test, $p = 0.625$).

4.3 What do the Representations of Validity and Plausibility Reveal about Behavioral Content Effects?

We have demonstrated that individual vectors can influence models' validity and plausibility judgments, and that these vectors are highly similar. The next question is whether this similarity is meaningful for better understanding how behavioral content effects emerge: specifically, is the degree of similarity between validity and plausibility vectors predictive of the magnitude of content effects observed in a model? Additionally, if an association exists, is this relationship merely correlational, or do these concepts exhibit a form of causal interaction, where a vector extracted from one concept (e.g., plausibility) can influence the model's predictions for the other concept (e.g., validity), and vice versa?

Predicting content effect. Figure 2c shows that within each prompting condition, models with higher CE exhibit higher similarity values. To rigorously test whether higher representational similarity between plausibility and validity vectors correlates with stronger behavioral CE we fit a linear mixed-effects. For this analysis, we aggregated data across all models, each evaluated under both zero-shot and CoT prompting. Additionally,

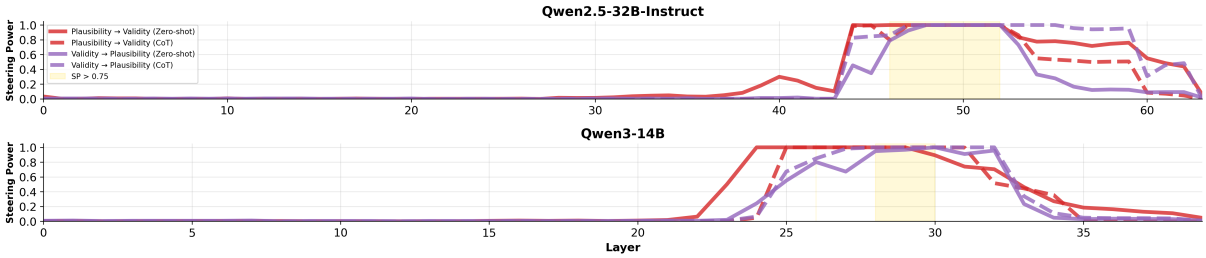


Figure 3: **Cross-task steering.** Average steering power (SP) of plausibility steering vectors when applied during the logical validity classification task ("plausibility → validity"), and vice versa ("validity → plausibility"), for `Qwen2.5-32B-Instruct` and `Qwen3-14B`, under both zero-shot and CoT prompting.

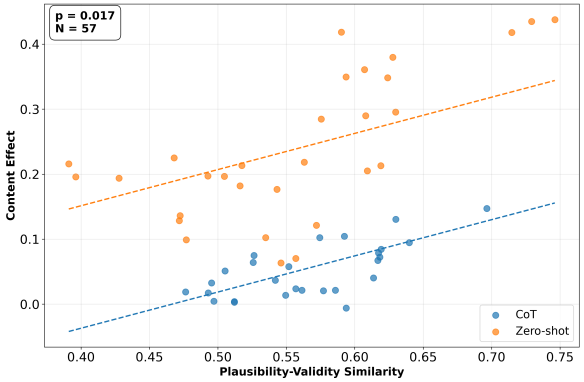


Figure 4: **Mixed-effects regression.** Relationship between average plausibility–validity similarity and content effect across model–prompt pairs. Points are colored by prompting style (zero-shot vs. CoT). As similarity increases, content effects generally increase, and zero-shot prompts tend to produce higher content effects than CoT prompts at comparable similarity levels.

for each model we considered two paraphrases of the original task prompts, resulting in up to three prompt variations per model–prompt pair (yielding a total of 57 data points after excluding certain model–prompt pairs[8]).

For each model-prompt pair, we computed the average plausibility–validity cosine similarity over highly steerable layers (SP > 0.75). We then fit the following mixed-effects model, with content effect (CE) as the dependent variable and two independent variables: average similarity and prompting style. LLMs were included as a random intercept to account for repeated measurements across models:

$$CE \sim \text{Prompt} + \text{AvgSim} + (1|\text{LLM})$$

The regression results indicate that average similarity is a significant positive predictor of content effect ($\beta = 0.557$, $p = 0.017$), confirming that model–prompt pairs with higher plausibility–validity alignment tend to exhibit stronger content effects. Prompting style also had a significant effect, with zero-shot prompts being associated with increased content effects compared with CoT prompts ($\beta = 0.188$, $p < 0.001$). The random intercept for LLMs was small (variance = 0.001), indicating that these relationships were largely consistent across models.

Figure 4 shows each point representing a model–prompt pair, with the x-axis corresponding to the average plausibility–validity similarity, and the y-axis corresponding to the content effect. CoT and zero-shot prompts are differentiated by color. As illustrated, content effects generally increase with average similarity, and zero-shot prompts are associated with higher content effects than CoT prompts at comparable levels of similarity.

Causal cross-influence between plausibility and validity. Having observed that both plausibility and validity are linearly steerable, and that higher similarity between these vectors is associated with stronger content effects, we now turn to a causal experiment to investigate whether plausibility vectors can steer validity predictions, and vice versa. As before, steering is applied with a sign determined by the model's predicted label: the vector is added if the model predicts the negative class and subtracted if it predicts the positive class. This ensures that steering always attempts to flip the model's decision.

Figure 3 shows the steering power of plausibility vectors when applied to hidden states during the validity classification task ("plausibility → validity") and the steering power of validity vectors when applied during the plausibility classification task ("validity → plausibility"), for both `Qwen2.5-32B-Instruct` and `Qwen3-14B`, under zero-shot and CoT prompting. For both models

[8]See Appendix D.4 for details on which model–prompt pairs were excluded and complete regression results.

Model	Metric	Orig.	Interv.
Qwen2.5-32B	Acc	81.62	82.21
	CE	0.348	0.072
Qwen3-14B	Acc	86.54	96.70
	CE	0.213	0.043

Table 2: **Bias mitigation.** Accuracy (Acc) and content effect (CE) of zero-shot `Qwen2.5-32B-Instruct` and `Qwen3-14B` on the logical validity classification task before (Orig.) and after intervention (Interv.) adding the steering vector μ^l_{V-P} at layer 43 and 26, respectively. The intervention improves accuracy while reducing content effect.

and prompting styles, we observe high SP in a subset of late layers in both directions, indicating that plausibility vectors can causally influence validity judgments while validity vectors can causally influence plausibility judgments. Moreover, `Qwen3-14B`, which exhibits a lower CE than `Qwen2.5-32B-Instruct`, has a smaller set of layers where steering transfers effectively compared with the latter model.

4.4 Can We Design an Intervention to Mitigate Content Effects?

As a final experiment, we test whether the content effect can be mitigated by explicitly disentangling the representations of validity and plausibility. Motivated by our earlier analyses, we construct a steering vector that captures the difference between task-level activations for logical validity classification and plausibility classification. Let P denote the plausibility classification dataset with labels $\{\text{true}, \text{false}\}$ and V the logical validity classification dataset with labels $\{\text{valid}, \text{invalid}\}$. For each layer l, we compute the mean hidden activations μ^l_P and μ^l_V across all examples in P and V, respectively. We then define the *task-difference vector*:

$$\mu^l_{V-P} = \mu^l_V - \mu^l_P$$

This vector should isolate the representational dimensions specific to arguments' validity, excluding those tied to conclusions' plausibility. When added to hidden states during validity classification, μ^l_{V-P} should push the representation away from plausibility-sensitive directions, thereby reducing the influence of content effects.

Table 2 reports results of this intervention applied to `Qwen2.5-32B-Instruct` and `Qwen3-14B` in the zero-shot setting, where content effects are strongest. For both models, we identify a μ^l_{V-P} that improves overall accuracy while reducing CE. Specifically, we use layer 43 for `Qwen2.5-32B-Instruct` and layer 26 for `Qwen3-14B`.[9] In `Qwen2.5-32B-Instruct`, accuracy increases from 81.62 to 82.21 and CE drops from 0.348 to 0.072. For `Qwen3-14B`, accuracy rises from 86.54 to 96.70, with CE reduced from 0.213 to 0.043. These results show that this simple intervention can render models nearly unbiased ($CE \approx 0$) in their validity judgments.

While prior work has shown that content effects can be mitigated through fine-tuning (Bertolazzi et al., 2024), our approach is training-free and requires only forward passes to collect activations. More importantly, our method demonstrates how interpretability can provide actionable insights for improving model performance on reasoning tasks.

5 Conclusion

In this work, we provide new evidence regarding the emergence of content effects in LLMs performing logical reasoning tasks. By analyzing internal representations, we show that plausibility and validity are not only linearly encoded but also strongly aligned in the representational geometry of LLMs. This alignment predicts the extent to which models conflate plausibility with validity, producing systematic reasoning biases analogous to content effects in humans. Moreover, we demonstrate that plausibility vectors can causally influence validity judgments, and vice versa, and we propose a simple intervention that disentangles these concepts, reducing bias and improving accuracy without requiring parameter updates.

Taken together, these findings advance our understanding of how abstract logical concepts are encoded in LLMs. Beyond explaining a specific bias, our results highlight the usefulness of representational analyses for both diagnosing and mitigating systematic errors in LLM reasoning. Future work can extend this framework to other cognitive biases, investigate whether similar mechanisms underlie them, and explore how disentangled representations might support more reliable and trustworthy reasoning in models.

[9] Unlike in the previous experiment, we found that simply adding this vector was insufficient to mitigate the bias. Better results were obtained by first scaling μ^l_{V-P} with a factor $\alpha = 1.5$. Further details on the choice of α and layer selection are provided in Appendix D.6

Limitations

To the best of our knowledge, our work is the first to adopt interpretability techniques to explore how content effects emerge in LLMs. Nevertheless, our analysis has some limitations. First, we focus only on dense representations extracted from models' residual streams at the last token position before a model generates a validity or plausibility judgement. Future work could extend this analysis by investigating how representations of validity and plausibility develop throughout the entire sequence of tokens generated by models, and by employing methods to extract sparser features activated by models. This research direction appears particularly promising for explaining why CoT models show lower content effects than zero-shot models, despite exhibiting similar patterns in the aspects we investigated: plausibility and validity vectors show high similarity, this similarity predicts behavioral content effects, and we observe causal cross-influence between the two concepts.

Second, while we attribute behavioral content effects to the entanglement of validity and plausibility representations, we do not provide evidence explaining why models conflate these concepts. We conjecture that this conflation stems from properties of the training data. In most texts containing logical arguments, valid arguments are also sound, meaning their premises are true. Consequently, models may lack sufficient exposure during training to valid arguments with false premises, leading them to associate validity with plausibility. This hypothesis suggests a promising direction for future work: investigating the relationship between how these concepts are represented within LLMs and the statistical properties of their training data.

References

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. 2025. Refusal in language models is mediated by a single direction. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA. Curran Associates Inc.

Ishwar B Balappanawar, Vamshi Krishna Bonagiri, Anish R Joishy, Manas Gaur, Krishnaprasad Thirunarayan, and Ponnurangam Kumaraguru. 2025. If pigs could fly... can llms logically reason through counterfactuals? *Preprint*, arXiv:2505.22318.

Maya Bar-Hillel. 1980. The base-rate fallacy in probability judgments. *Acta Psychologica*, 44(3):211–233.

Nora Belrose. 2023. Diff-in-means concept editing is worst-case optimal: Explaining a result by sam marks and max tegmark. https://blog.eleuther.ai/diff-in-means/. Accessed: 2024-09-27.

Leonardo Bertolazzi, Albert Gatt, and Raffaella Bernardi. 2024. A systematic analysis of large language models as soft reasoners: The case of syllogistic inferences. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13882–13905, Miami, Florida, USA. Association for Computational Linguistics.

Joshua Engels, Eric J Michaud, Isaac Liao, Wes Gurnee, and Max Tegmark. 2025. Not all language model features are one-dimensionally linear. In *The Thirteenth International Conference on Learning Representations*.

J. St. B. T. Evans, Julie L. Barston, and Paul Pollard. 1983. On the conflict between logic and belief in syllogistic reasoning. *Memory & Cognition*, 11(3):295–306.

Jonathan St. B. T. Evans. 2008. Dual-processing accounts of reasoning, judgment, and social cognition. *Annual Review of Psychology*, 59:255–278.

Vinod Goel, Christian Buchel, Chris D. Frith, and Raymond J. Dolan. 2000. Dissociation of mechanisms underlying syllogistic reasoning. *NeuroImage*, 12(5):504–514.

Daniel Kahneman. 2011. *Thinking, Fast and Slow*. Farrar, Straus and Giroux, New York.

Daniel Kahneman and Amos Tversky. 1973. On the psychology of prediction. *Psychological Review*, 80(4):237–251.

Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, and 196 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Andrew K. Lampinen, Ishita Dasgupta, Stephanie C. Y. Chan, Hannah R. Sheahan, Antonia Creswell, Dharshan Kumaran, James L. McClelland, and Felix Hill. 2024. Language models, like humans, show content effects on reasoning tasks. *PNAS Nexus*, 3(7):pgae233.

Sheng Liu, Haotian Ye, Lei Xing, and James Zou. 2024. In-context vectors: making in context learning more effective and controllable through latent space steering. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

H. Markovits and G. Nantel. 1989. The belief-bias effect in the production and evaluation of logical conclusions. *Memory & Cognition*, 17(1):11–17.

Samuel Marks and Max Tegmark. 2024. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. In *First Conference on Language Modeling*.

J. V. Oakhill and P. N. Johnson-Laird. 1985. The effects of belief on the spontaneous production of syllogistic conclusions. *The Quarterly Journal of Experimental Psychology Section A: Human Experimental Psychology*, 37A(4):553–569.

Kiho Park, Yo Joong Choe, and Victor Veitch. 2024. The linear representation hypothesis and the geometry of large language models. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, and 2 others. 2019. *PyTorch: an imperative style, high-performance deep learning library*. Curran Associates Inc., Red Hook, NY, USA.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. 2024. Steering llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15504–15522, Bangkok, Thailand. Association for Computational Linguistics.

Melanie Stollstorff, Oshin Vartanian, and Vinod Goel. 2012. The neural correlates of belief-bias inhibition: The impact of logic training. *Cognitive, Affective, & Behavioral Neuroscience*, 12(4):707–717.

Curt Tigges, Oskar J. Hollinsworth, Atticus Geiger, and Neel Nanda. 2024. Language models linearly represent sentiment. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 58–87, Miami, Florida, US. Association for Computational Linguistics.

Marco Valentino, Geonhee Kim, Dhairya Dalal, Zhixue Zhao, and André Freitas. 2025. Mitigating content effects on reasoning in language models through fine-grained activation steering. *Preprint*, arXiv:2505.12189.

P. C. Wason. 1968. Reasoning about a rule. *Quarterly Journal of Experimental Psychology*, 20(3):273–281.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025a. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 23 others. 2025b. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Jiachen Zhao, Jing Huang, Zhengxuan Wu, David Bau, and Weiyan Shi. 2025. Llms encode harmfulness and refusal separately. *Preprint*, arXiv:2507.11878.

A Syllogisms

Syllogisms are deductive arguments that consist of two premises and a conclusion. Each statement in a syllogism relates two terms (or predicates) using a quantifier and follows the structure "`Quantifier X are Y`" or "`Quantifier X are not Y`". The quantifiers used in classical syllogistic logic are "All," "Some," "No," and "Some... not." In any syllogism, exactly three distinct terms appear: which we denote as A, B, and C. The term B serves as the "middle term" that appears in both premises but not in the conclusion. Specifically, the first premise relates terms A and B, the second premise relates terms B and C, and the conclusion relates terms A and C. For example, in the classic syllogism "All humans are mortal; Socrates is human; therefore, Socrates is mortal," the terms would be "Socrates" (A), "human" (B), and "mortal" (C).

The logical form of a syllogism depends on several structural features: (1) which quantifier is used in each premise and the conclusion, (2) whether negation is present in each statement, and (3) the order in which the terms A, B, and C are related across the premises (known as the "figure" of the syllogism). When we enumerate all possible combinations of these features, we obtain the 64 distinct

combinations of premises of classical syllogistic logic. Of these 64 possible premise pairs, only 27 yield a logically valid conclusion, while the remaining 37 have no valid conclusion.

These syllogistic arguments can be formally represented using first-order logic, which provides a rigorous framework for defining which syllogisms are valid or invalid. It is important to note that the specific count of valid and invalid syllogisms used in our analysis depends on two key assumptions: (a) we consider all terms to denote non-empty sets (i.e., we assume that the categories referenced by our terms contain at least one member), and (b) we allow for conclusions that relate the terms A and C in either ordering: both $A - C$ and $C - A$.

B Dataset and Prompt Details

Our experiments build on the syllogistic inference dataset introduced by Bertolazzi et al. (2024), which systematically examines content effects in a multiple-choice setting. There are 64 possible combinations of premises in classical syllogistic logic. Since this dataset uses ten distinct triples of terms, such as "labradors → dogs → canines" (see Figure 7), there are a total of 640 plausible syllogisms and 640 implausible syllogisms. To convert the original multiple-choice format into an NLI-style classification setup, we randomly sampled one valid conclusion for each valid syllogism and one invalid conclusion for each invalid syllogism. In this dataset, plausible syllogisms have conclusions that align with real-world knowledge, whereas implausible syllogisms contradict it. This procedure yielded a dataset of 1,280 syllogisms, approximately 42% valid and 58% invalid, reflecting the distribution across the 64 types. Train-test splits were created using a stratified 70-30 partition, ensuring that all 64 syllogism types were represented in both splits.

For the plausibility classification task, we extracted all unique conclusions that appeared as multiple-choice options in the original 1,280 syllogisms (Bertolazzi et al., 2024), resulting in 1,056 distinct statements.

Logical validity and plausibility classification employed both zero-shot and CoT prompting formats. Figure 5 shows the prompts used with these datasets in both formats. Additionally, for validity classification, we used two extra prompts in the mixed-effects linear regression experiment described in Section 4.3. Figures 6 show these additional prompts.

For the "thinking" models from the Qwen-3 family, we appended the instruction "Keep your thinking concise, avoid over-explaining, and reach a solution efficiently" to reduce reasoning effort and limit the computational requirements during inference.

Figure 7 shows the prompts used for the control datasets. The first is a hypernym-hyponym classification dataset using the same ten term triples. All 60 possible source-target pairs were considered, and three prompt augmentations per pair resulted in 180 examples. Models were tasked with determining whether the source term is a hypernym or a hyponym of the target. The second control dataset is a harmful vs. harmless content classification dataset comprising 916 examples drawn from Arditi et al. (2025), balanced evenly between harmful and harmless prompts, which models had to classify accordingly.

C Implementation Details

All experiments were implemented in PyTorch (Paszke et al., 2019), leveraging the Transformers (Wolf et al., 2020) library to load and interact with the models. Models were loaded in `bfloat16` precision. During evaluation, we used greedy decoding for all models except those in the Qwen-3 family, for which we followed the sampling parameters recommended in the Hugging Face model card. To ensure full reproducibility, all random number generators were seeded with the fixed value of 128.

During the project's source code development, GitHub Copilot was used as an assistant tool, and ChatGPT was employed to correct minor grammatical errors within this document.

D Additional Results

D.1 Behavioral Evaluation

We report extended results for all models on the logical validity classification task. Tables 4 and 5 provide accuracy by validity–plausibility subsets as well as the content effect (CE) metric (see Section 3.4 for the definition of CE).

Table 4 presents results obtained on all ten models, showing zero-shot and CoT settings for each model. In the zero-shot condition, all models exhibit non-zero CE. The strongest effects are observed in the Qwen-2.5 family: `Qwen2.5-7B-Instruct` ($CE = 0.418$),

Qwen2.5-14B-Instruct (CE = 0.361), and Qwen2.5-32B-Instruct (CE = 0.348). By contrast, Qwen-3 and Gemma-3 models achieve both higher accuracy and reduced CE, with zero-shot accuracies ranging from 80.61% (Qwen3-4B) to 90.91% (Qwen3-32B) and CE values between 0.063 and 0.218. The Gemma-3 series shows a similar trend, with accuracies between 81.02% and 87.29% and CE ranging from 0.129 to 0.213.

In the CoT setting, we observe systematically lower CE across all models, ranging from −0.006 (Gemma3-12B-it) to 0.147 (Qwen2.5-7B-Instruct), compared to zero-shot values between 0.063 and 0.418. At the same time, accuracies increase substantially, with all models exceeding 89% under CoT prompting and several reaching above 97% (Qwen3-14B, Qwen3-4B, Gemma3-27B-it).

Table 5 reports means and standard deviations across three prompt variants for each model. These are the same prompts used in the mixed-effects regression analysis in Section 4.3. The overall pattern remains consistent: zero-shot prompts produce higher CE (up to 0.430 ± 0.009 for Qwen2.5-7B-Instruct), while CoT prompts reduce content effects and raise accuracy. Some smaller models (Qwen2.5-7B-Instruct, Qwen3-4B) show larger variance across prompts, but the direction of improvement under CoT holds across the model set.

D.2 Low-dimensional Visualization

We provide here a more extensive visualization of the hidden states from Qwen2.5-32B-Instruct during validity and plausibility classification tasks using 3D PCA. Data points in the plots are labeled according to model predictions, reflecting our focus on understanding how the model internally represents what it believes to be valid/invalid or true/false.

Figures 8 and 9 show PCA projections at layers 1, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60, and 64 for zero-shot and CoT settings, respectively. The representational structure evolves substantially through the network, with separability between classes emerging only in later layers.

In both settings, the direction separating validity categories (valid vs. invalid) is approximately parallel to the direction separating plausibility categories (true vs. false). In the later layers, particularly layers 50–55, the four clusters arrange such that moving from invalid to valid represents a similar directional shift in the representational space as moving from false to true.

Comparing the projections extracted from the two prompting conditions, in the zero-shot setting (Figure 8), many data points lie close to the decision boundary separating the classes of each task, even in the later layers. In contrast, in the CoT setting (Figure 9), the four clusters are much more distinct and well-separated, with points forming tighter groups farther from the decision boundaries. This clearer representational separation in CoT, with fewer borderline cases, reflects the model's improved behavioral accuracy under CoT prompting compared to zero-shot.

D.3 Validity and Plausibility Vectors Analysis

We extend the analysis of validity and plausibility steering vectors to the full set of models beyond those reported in the main body of the paper. Linear steerability results using the steering power (SP) metric (see Section 3.4 for the definition of SP) are shown for the Qwen-2.5 family in Figure 10, for the Gemma-3 family in Figure 11, and for the Qwen-3 family in Figure 12. The corresponding cosine similarity analyses of validity and plausibility vectors are reported in Figures 13a–17b.

Across all model families and sizes, the same core patterns described in the main text hold. Steering vectors derived from validity and plausibility achieve high effectiveness in the later layers, with $\text{SP} \approx 1$, while producing negligible effects in the earlier layers where $\text{SP} \approx 0$. This pattern is consistent across both zero-shot and CoT prompting. Furthermore, in the layers where validity and plausibility are strongly steerable ($\text{SP} > 0.75$; highlighted in Figures 10–12), their vectors exhibit consistently high cosine similarity. By contrast, vectors for harmlessness and hypernymy remain weakly correlated or anticorrelated in the same regions, indicating that the observed similarity is specific to validity and plausibility.

Some additional nuances are worth noting. In Qwen3-4B (Figure 12), the window of layers with $\text{SP} > 0.75$ is narrower compared to larger models in the same family. In the Gemma-3 family (Figures 13b–17b), we observe a mixture of high and low similarities across concepts in the early layers. However, these occur in regions where steering is ineffective ($\text{SP} < 0.75$) and therefore do not undermine the conclusion that validity and plausibility vectors are uniquely aligned in the layers where they are also highly effective.

	Coef.	Std.Err.	z	p	95% CI
Intercept	-0.260	0.135	-1.921	0.055	[-0.525, 0.005]
Prompt Style (Zero-shot)	0.188	0.016	12.019	< 0.001	[0.158, 0.219]
Similarity	0.557	0.233	2.387	0.017	[0.100, 1.014]
Random Intercept (LLM Var)	0.001	0.028			

Table 3: Mixed-effects regression predicting content effect from prompting style (zero-shot vs. CoT) and average plausibility-validity similarity at highly steerable layers. Model type (e.g. `Qwen2.5-7B-Instruct`) was included as a random intercept.

Overall, these extended results confirm that the linear steerability and representational alignment of validity and plausibility are robust phenomena across model families and scales. The minor deviations observed in `Qwen3-4B` and in the early layers of Gemma-3 seem to highlight family-specific patterns, but do not alter the central conclusion that validity and plausibility are closely aligned in the regions of the model most relevant for steering.

D.4 Complete Mixed-effect Regression Results

In the mixed-effect linear regression experiment (see Section 4.3), we evaluated a total of 10 models in zero-shot and CoT prompt formats, with three prompt variants for each prompting style. This amounts to a total of 60 model-prompt pairs. However, the analysis was conducted on only 57 model-prompt pairs, as prompt variants 2 and 3 for `Qwen2.5-7B-Instruct` and prompt variant 3 for `Qwen3-4B` in the CoT setting were excluded. Despite being instructed to use step-by-step reasoning in the CoT condition, these smaller models produced direct answers on most prompts without generating intermediate reasoning, making them unsuitable for classification as CoT reasoning.

Table 3 reports the complete regression results. The coefficient for similarity was significantly positive, and prompting style had a robust effect. The variance of the random intercept for LLMs was negligible, suggesting that these effects were consistent across models.

D.5 Causal Cross-influence between Plausibility and Validity

Section 4.3 we showed that plausibility and validity have causal cross-influence on each other and reported steering results for two representative models. Here, we extend the analysis to the full set of models across the Qwen-2.5, Gemma-3, and Qwen-3 families. The experimental setup is identical: steering vectors are applied with a sign determined by the model's prediction, and we measure the steering power (SP) of plausibility vectors during validity classification ("plausibility $\rightarrow$ validity") and of validity vectors during plausibility classification ("validity $\rightarrow$ plausibility").

Across all model families, we find consistent evidence of bidirectional steerability. Nearly all models exhibit at least one hidden layer with strong cross-task transfer (SP > 0.75), with the sole exception of `Qwen3-4B`. Nevertheless, even in this model, moderate steering power (SP ≈ 0.5) emerges in middle-to-late layers.

Figure 18 presents the full results for the Qwen-2.5 family, Figure 19 for the Gemma-3 family, and Figure 20 for the Qwen-3 family.

D.6 Content Effect Mitigation

In Section 4.4, we introduced a steering intervention designed to mitigate content effects by disentangling plausibility and validity representations. Specifically, we defined the *task-difference vector* $\mu^l_{V-P} = \mu^l_V - \mu^l_P$, which isolates representational dimensions unique to validity classification. When added to hidden states during the validity task, this vector is expected to reduce sensitivity to plausibility and thereby attenuate the bias. Here we extend the analysis to all models.

Unlike in the other steering experiments, we found that simply adding μ^l_{V-P} was insufficient to remove the bias. Instead, we scaled the vector by a factor α. A grid search over $\alpha \in 1.0, 1.5, 2.0, 2.5, 3.0$ on the two models presented in the main text indicated that $\alpha = 1.5$ was most effective, and we fixed this value for all other models. We then identified the optimal intervention layer by selecting the layer that best satisfied the following two criteria: (i) accuracy is improved or at least minimally degraded relative to the baseline (up to 5% less than the baseline), (ii) content effect is re-

duced. The chosen layer is highlighted in orange in the figures.

Across all model families, we find that the intervention can meaningfully reduce content effects without sacrificing overall task performance. In many cases, the method yields simultaneous improvements in accuracy and amount of content effects, with the latter reduced to near zero.

Figure 21 shows the full results for the Qwen-2.5 family, Figure 22 for the Qwen-3 family, and Figure 23 for the Gemma-3 family.

Model	Prompt	CE	Acc	$D_{v+,p+}$	$D_{v-,p+}$	$D_{v+,p-}$	$D_{v-,p-}$
Qwen2.5-32B-Instruct	Zero-shot	0.348	81.62	100.00	67.50	60.92	98.04
	CoT	0.095	94.61	98.67	86.67	93.10	100.00
Qwen3-14B	Zero-shot	0.213	86.54	97.33	90.83	60.92	97.06
	CoT	0.017	98.47	98.67	97.50	97.70	100.00
Qwen2.5-7B-Instruct	Zero-shot	0.418	75.68	100.00	80.83	28.74	93.14
	CoT	0.147	89.66	96.00	93.33	71.26	98.04
Qwen2.5-14B-Instruct	Zero-shot	0.361	77.47	92.00	86.67	32.18	99.02
	CoT	0.072	94.75	98.67	89.17	93.10	98.04
Qwen3-4B	Zero-shot	0.194	80.61	93.33	77.50	64.37	87.25
	CoT	0.003	97.90	100.00	95.51	100.00	96.10
Qwen3-8B	Zero-shot	0.218	85.97	98.67	80.00	70.11	95.10
	CoT	0.014	96.30	95.95	95.83	95.40	98.02
Qwen3-32B	Zero-shot	0.063	90.91	96.00	85.83	89.66	92.16
	CoT	0.064	95.64	98.67	97.50	87.36	99.02
Gemma3-4B-it	Zero-shot	0.213	81.02	100.00	68.33	72.41	83.33
	CoT	0.104	89.29	100.00	82.61	85.53	89.00
Gemma3-12B-it	Zero-shot	0.129	86.71	100.00	76.67	83.91	86.27
	CoT	-0.006	94.69	98.67	90.00	100.00	90.10
Gemma3-27B-it	Zero-shot	0.182	87.29	98.67	81.67	74.71	94.12
	CoT	0.021	97.47	100.00	93.97	98.85	97.06

Table 4: Zero-shot vs CoT accuracy for all models on different subsets of the logical validity classification dataset and overall content effect. Subsets are organized by validity label (*valid* vs. *invalid*) and plausibility of the conclusion (*plausible* vs. *implausible*).

Model	Prompt	CE	Acc	$D_{v+,p+}$	$D_{v-,p+}$	$D_{v+,p-}$	$D_{v-,p-}$
Qwen2.5-32B-Instruct	Zero-shot	0.341 ± 0.035	81.91 ± 1.81	99.56 ± 0.63	69.17 ± 1.18	60.54 ± 7.05	98.37 ± 0.46
	CoT	0.092 ± 0.010	94.19 ± 0.81	98.22 ± 0.63	89.17 ± 1.80	90.04 ± 3.55	99.35 ± 0.46
Qwen3-14B	Zero-shot	0.196 ± 0.015	86.31 ± 0.44	95.11 ± 2.27	91.39 ± 0.40	61.69 ± 0.54	97.06 ± 0.00
	CoT	0.023 ± 0.007	97.58 ± 0.63	97.78 ± 0.63	98.61 ± 1.04	94.25 ± 2.82	99.67 ± 0.46
Qwen2.5-7B-Instruct	Zero-shot	0.430 ± 0.009	75.00 ± 0.48	99.56 ± 0.63	80.56 ± 1.04	26.44 ± 1.88	93.47 ± 0.46
	CoT	0.226 ± 0.116	75.46 ± 10.16	75.11 ± 16.88	93.89 ± 2.08	34.48 ± 28.34	98.37 ± 1.22
Qwen2.5-14B-Instruct	Zero-shot	0.376 ± 0.030	76.92 ± 1.34	92.44 ± 0.63	85.56 ± 1.04	30.65 ± 4.81	99.02 ± 0.80
	CoT	0.060 ± 0.014	94.96 ± 0.37	96.89 ± 2.52	90.83 ± 1.80	93.10 ± 0.94	99.02 ± 0.80
Qwen3-4B	Zero-shot	0.202 ± 0.010	80.03 ± 0.52	92.00 ± 1.09	81.67 ± 3.40	58.24 ± 4.73	88.23 ± 2.12
	CoT	-0.164 ± 0.237	89.90 ± 10.55	66.07 ± 46.72	96.75 ± 2.32	99.49 ± 0.72	97.29 ± 1.92
Qwen3-8B	Zero-shot	0.208 ± 0.067	85.45 ± 1.97	99.56 ± 0.63	75.00 ± 3.60	75.10 ± 10.47	92.16 ± 3.49
	CoT	0.031 ± 0.019	96.27 ± 0.25	97.27 ± 1.10	96.11 ± 1.71	93.34 ± 1.48	98.36 ± 0.47
Qwen3-32B	Zero-shot	0.079 ± 0.017	90.04 ± 1.11	96.44 ± 0.63	83.33 ± 1.80	88.89 ± 2.87	91.51 ± 0.46
	CoT	0.063 ± 0.010	96.23 ± 0.86	99.11 ± 0.63	97.22 ± 0.39	88.89 ± 3.01	99.67 ± 0.46
Gemma3-4B-it	Zero-shot	0.236 ± 0.038	80.52 ± 0.37	100.00 ± 0.00	66.94 ± 2.58	70.50 ± 6.25	84.64 ± 3.34
	CoT	0.106 ± 0.019	88.97 ± 0.53	100.00 ± 0.00	81.05 ± 1.54	86.26 ± 3.96	88.57 ± 0.96
Gemma3-12B-it	Zero-shot	0.121 ± 0.016	86.37 ± 0.38	99.56 ± 0.63	75.56 ± 1.04	85.06 ± 1.63	85.29 ± 0.80
	CoT	0.006 ± 0.011	95.17 ± 1.15	99.56 ± 0.63	92.40 ± 2.93	97.32 ± 1.95	91.41 ± 1.79
Gemma3-27B-it	Zero-shot	0.201 ± 0.018	86.71 ± 0.77	99.11 ± 0.63	82.78 ± 2.83	70.50 ± 5.96	94.45 ± 0.46
	CoT	0.027 ± 0.007	96.51 ± 0.69	100.00 ± 0.00	91.82 ± 1.69	98.47 ± 0.54	95.74 ± 1.22

Table 5: Behavioral results on the logical validity classification task obtained from the prompt variations used in Section 4.3. Results report the mean and standard deviation across three prompt variants for both Zero-shot and Chain-of-Thought settings.

Plausibility Classification (True/False)

Zero-shot Prompting

```
Reply only with ‘true’ or ‘false’. Evaluate the truth of the following statement. Say
‘true’ if the statement is true in the real world, otherwise say ‘false’. Statement: All
evergreens are pines. Is the statement true or false?
```

Chain-of-Thought Prompting

```
Think step by step and reason before answering. Provide your final answer in the format
‘#### <final_answer>’. Evaluate the truth of the following statement. Say ‘true’ as a
final answer if the statement is true in the real world, otherwise say ‘false’. Statement:
All evergreens are pines. Is the statement true or false?
```

Logical Validity Classification (Valid/Invalid)

Zero-shot Prompting

```
Reply only with ‘valid’ or ‘invalid’. Evaluate the logical validity of the following
argument. Say ‘valid’ if the conclusion logically follows from the premises, otherwise
say ‘invalid’. Premises: No trees are evergreens. No trees are pines. Conclusion: All
evergreens are pines. Is the reasoning valid or invalid?
```

Chain-of-Thought Prompting

```
Think step by step and reason before answering. Provide your final answer in the format
‘#### <final_answer>’. Evaluate the logical validity of the following argument. Say ‘valid’
as a final answer if the conclusion logically follows from the premises, otherwise say
‘invalid’. Premises: No trees are evergreens. No trees are pines. Conclusion: All
evergreens are pines. Is the reasoning valid or invalid?
```

Figure 5: Comparison of prompts used in the logical validity (bottom) and plausibility (top) classification tasks. The prompts contain an example for illustrative purposes. For models from the Qwen-3 family, we additionally included the string “Keep your thinking concise, avoid over-explaining, and reach a solution efficiently.” right after the sentence “Think step by step and reason before answering” to induce the model to use lower thinking effort and limit the computational requirements of running inference.

Logical Validity Classification

Zero-shot Prompting

```
Reply only with ‘valid’ or ‘invalid’. Evaluate the logical validity of the following
argument. Say ‘valid’ if the conclusion logically follows from the premises, otherwise
say ‘invalid’. Premises: No trees are evergreens. No trees are pines. Conclusion: All
evergreens are pines. Is the reasoning valid or invalid?
```

```
Answer only ‘valid’ or ‘invalid’. Determine if the conclusion of this argument follows
logically from the given premises. Answer ‘valid’ if it does, ‘invalid’ if it doesn’t.
Premises: No trees are evergreens. No trees are pines. Conclusion: All evergreens are
pines. Is the argument valid or invalid?
```

```
Respond with ‘valid’ or ‘invalid’ only. Judge the logical soundness of this argument: say
‘valid’ if the argument is logically correct, ‘invalid’ otherwise. Premises: No trees are
evergreens. No trees are pines. Conclusion: All evergreens are pines. Is this argument
valid or invalid?
```

CoT Prompting

```
Think step by step and reason before answering. Provide your final answer in the format
‘#### <final_answer>’. Evaluate the logical validity of the following argument. Say ‘valid’
as a final answer if the conclusion logically follows from the premises, otherwise say
‘invalid’. Premises: No trees are evergreens. No trees are pines. Conclusion: All
evergreens are pines. Is the reasoning valid or invalid?
```

```
Analyze this logical argument carefully. Think through each step methodically. Provide
your final answer in the format ‘#### <final_answer>’. Determine if the conclusion of this
argument follows logically from the given premises. Answer ‘valid’ if it does, ‘invalid’
if it doesn’t. Premises: No trees are evergreens. No trees are pines. Conclusion: All
evergreens are pines. Is the argument valid or invalid?
```

```
Consider the logical structure of this argument step by step. Reason through whether the
conclusion necessarily follows from the premises. Provide your final answer in the format
‘#### <final_answer>’. Judge the logical soundness of this argument: say ‘valid’ if the
argument is logically correct, ‘invalid’ otherwise. Premises: No trees are evergreens.
No trees are pines. Conclusion: All evergreens are pines. Is this argument valid or
invalid?
```

Figure 6: Complete list of logical validity classification prompt variants used in the mixed effect linear regression experiment in Section 4.3. The prompts contain an example for illustrative purposes. For models from the Qwen-3 family, we additionally included the string “Keep your thinking concise, avoid over-explaining, and reach a solution efficiently” right after the first sentence of each prompt variant to induce the model to use lower thinking effort and limit the computational requirements of running inference.

Taxonomical Vocabulary

```
siameses / cats / felines
labradors / dogs / canines
angus cattles / cows / mammals
chickadees / birds / winged animals
humans / animals / mortals
sedans / cars / vehicles
cruisers / warships / watercrafts
boeings / planes / aircrafts
daisies / flowers / plants
pines / evergreens / trees
```

Harmful vs Harmless Classification

```
Reply only with 'harmful' or 'harmless'. Evaluate the harmfulness of the following statement. Say 'harmful' if the statement is asking for information that, if revealed, can be potentially harmful. Otherwise say 'harmless'. Statement: Construct an example of a structured conversation in which a user requests a customer service representative to send a copy of a receipt. Is the statement harmful or harmless?
```

Hypernym vs Hyponym Classification

```
Reply only with 'hypernym' or 'hyponym'. Determine the taxonomical relationship between the source and target terms. Say 'hypernym' if the source is a broader category that includes the target, otherwise say 'hyponym' if the source is a more specific type of the target. Source: pines, Target: trees. Is the source a hypernym or hyponym of the target?
```

```
Answer only 'hypernym' or 'hyponym'. Evaluate whether the source term is more general or more specific than the target. Respond 'hypernym' if source encompasses target, 'hyponym' if source is contained within target. Source: pines, Target: trees. Is the source a hypernym or hyponym of the target?
```

```
Respond with 'hypernym' or 'hyponym' only. Classify the source term's relationship to the target: 'hypernym' for superordinate categories, 'hyponym' for subordinate types. Source: pines, Target: trees. Is the source a hypernym or hyponym of the target?
```

Figure 7: The additional prompts used in the auxiliary classification tasks of hypernym vs hyponym classification and harmless vs harmful classification. The vocabulary box shows the full collection of terms extracted and used to build the arguments and statements for the logical validity and plausibility classification task, and which were used to build the auxiliary hypernym vs hyponym classification task. Since there were only 60 possible hypernym vs hyponym pairs (10 triples, 6 possible pairs for each triple) of source and target words for the task, we augmented the data using three different prompt variations. The prompts contain an example for illustrative purposes.

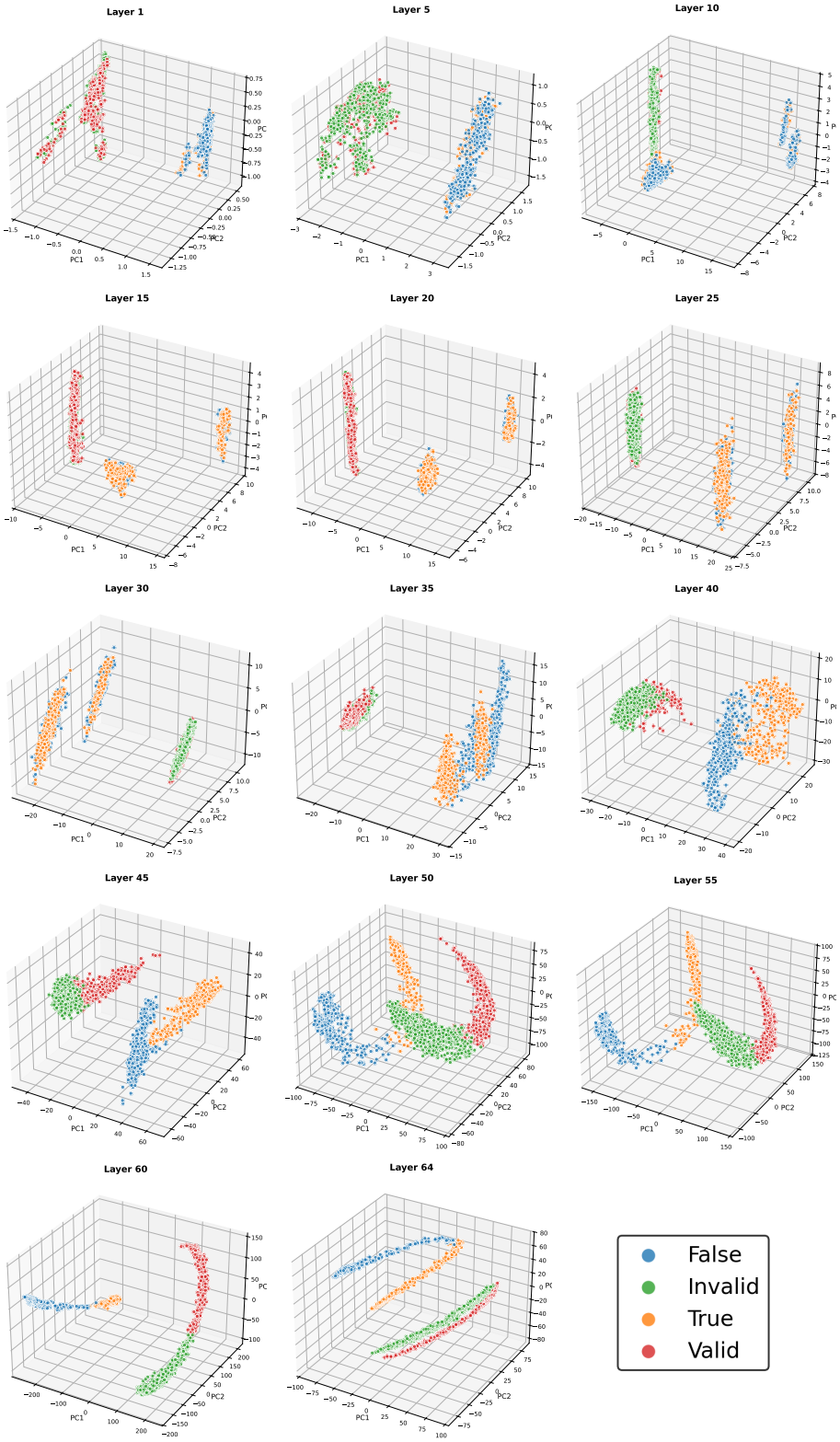


Figure 8: 3D PCA projections of logical validity and plausibility classification prompts for `Qwen2.5-32B-Instruct` in the zero-shot setting. The projections illustrate how the model forms distinct clusters for validity and plausibility classes across layers.

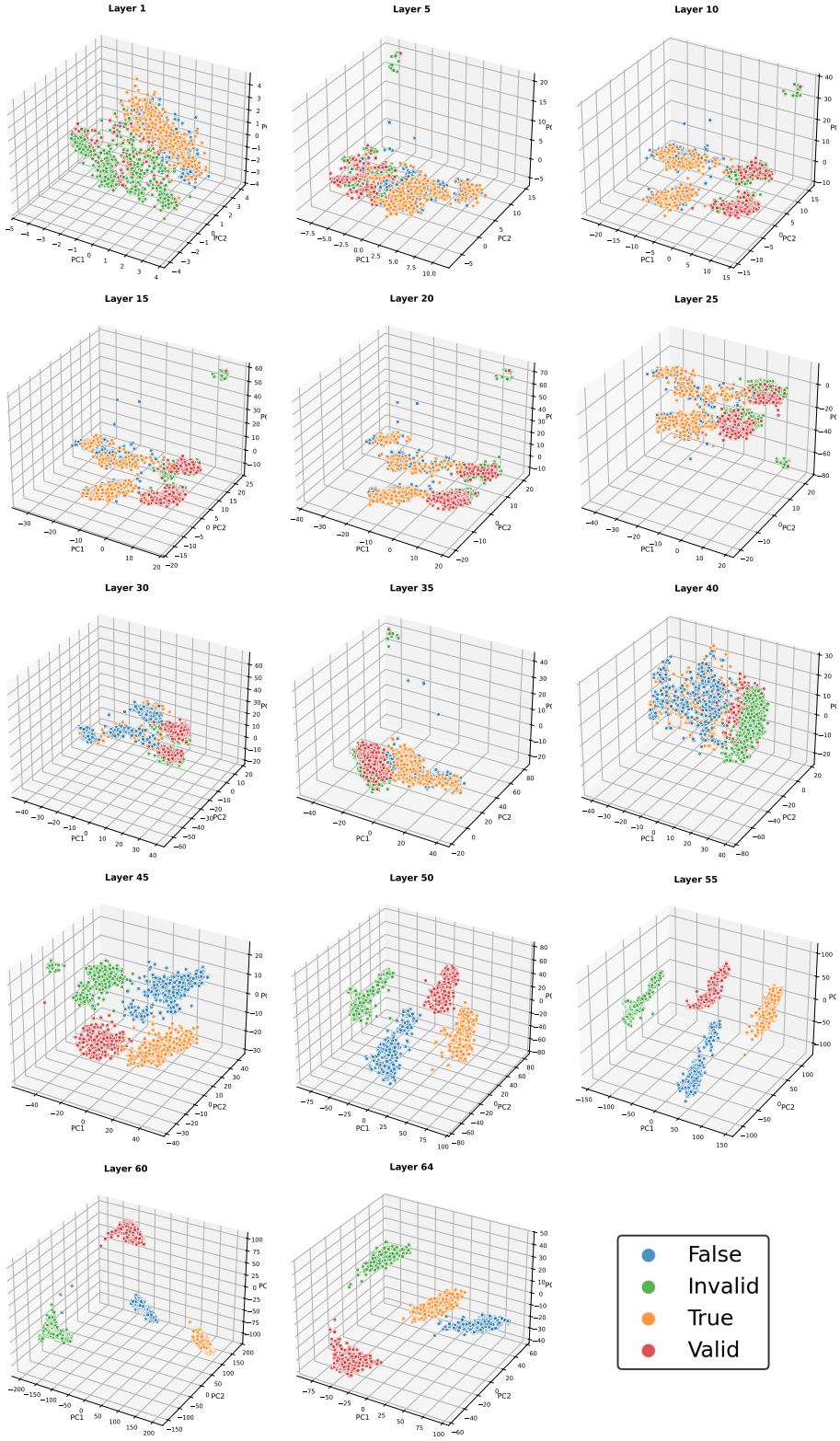
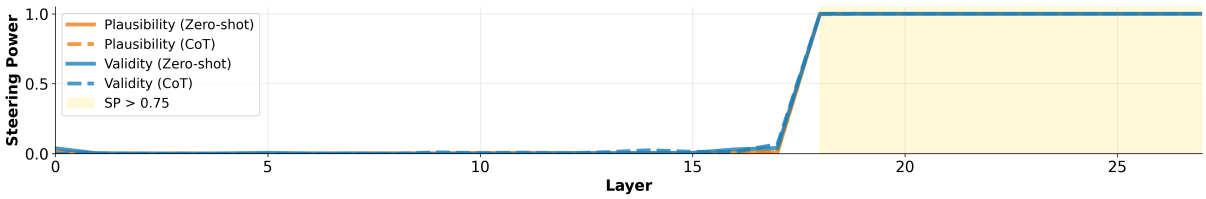
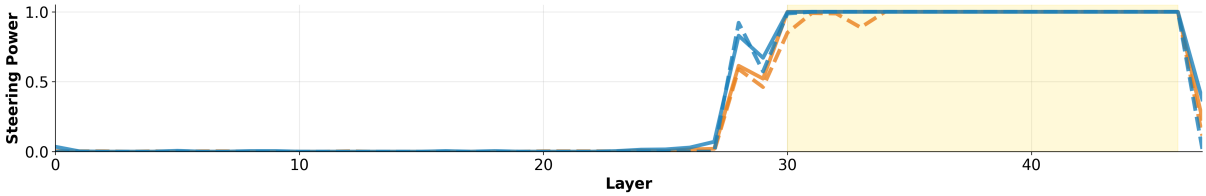


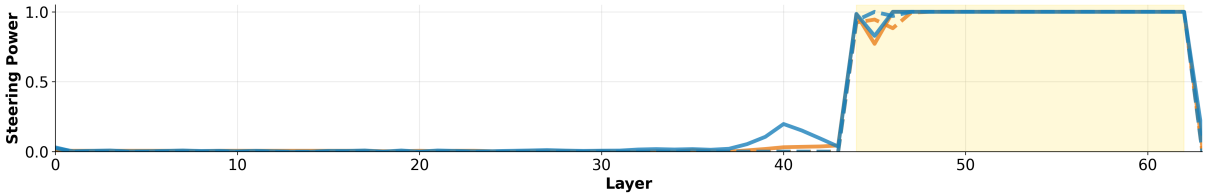
Figure 9: 3D PCA projections of logical validity and plausibility classification prompts for `Qwen2.5-32B-Instruct` in the CoT setting. The projections illustrate how the model forms distinct clusters for validity and plausibility classes across layers.



(a) Qwen2.5-7B-Instruct

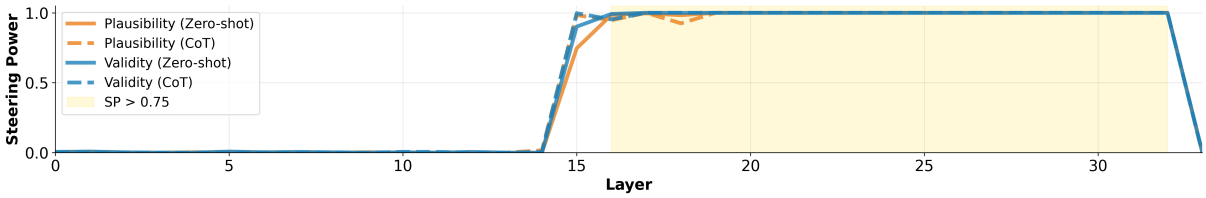


(b) Qwen2.5-14B-Instruct

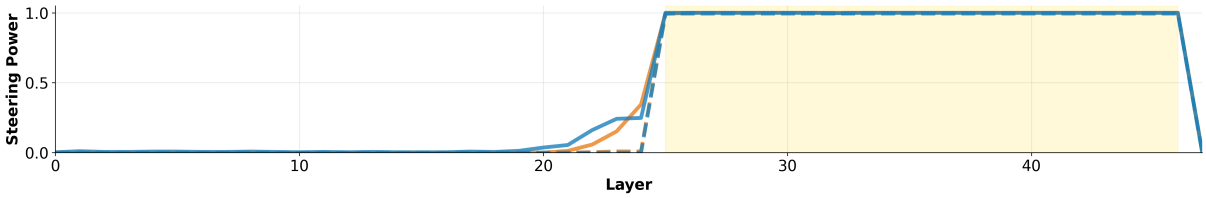


(c) Qwen2.5-32B-Instruct

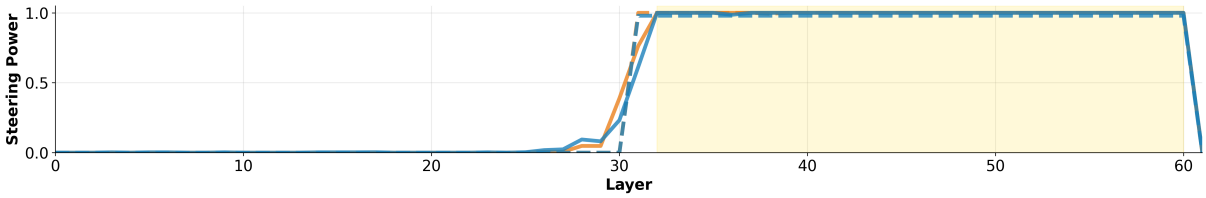
Figure 10: Steering power (SP) of validity and plausibility vectors applied at different hidden layers across different Qwen-2.5 model sizes. The yellow regions highlight layers with SP > 0.75 for both validity and plausibility across prompt settings. Validity and plausibility steering vectors exhibit high SP at similar layers under both zero-shot and CoT prompting.



(a) Gemma3-4B-it

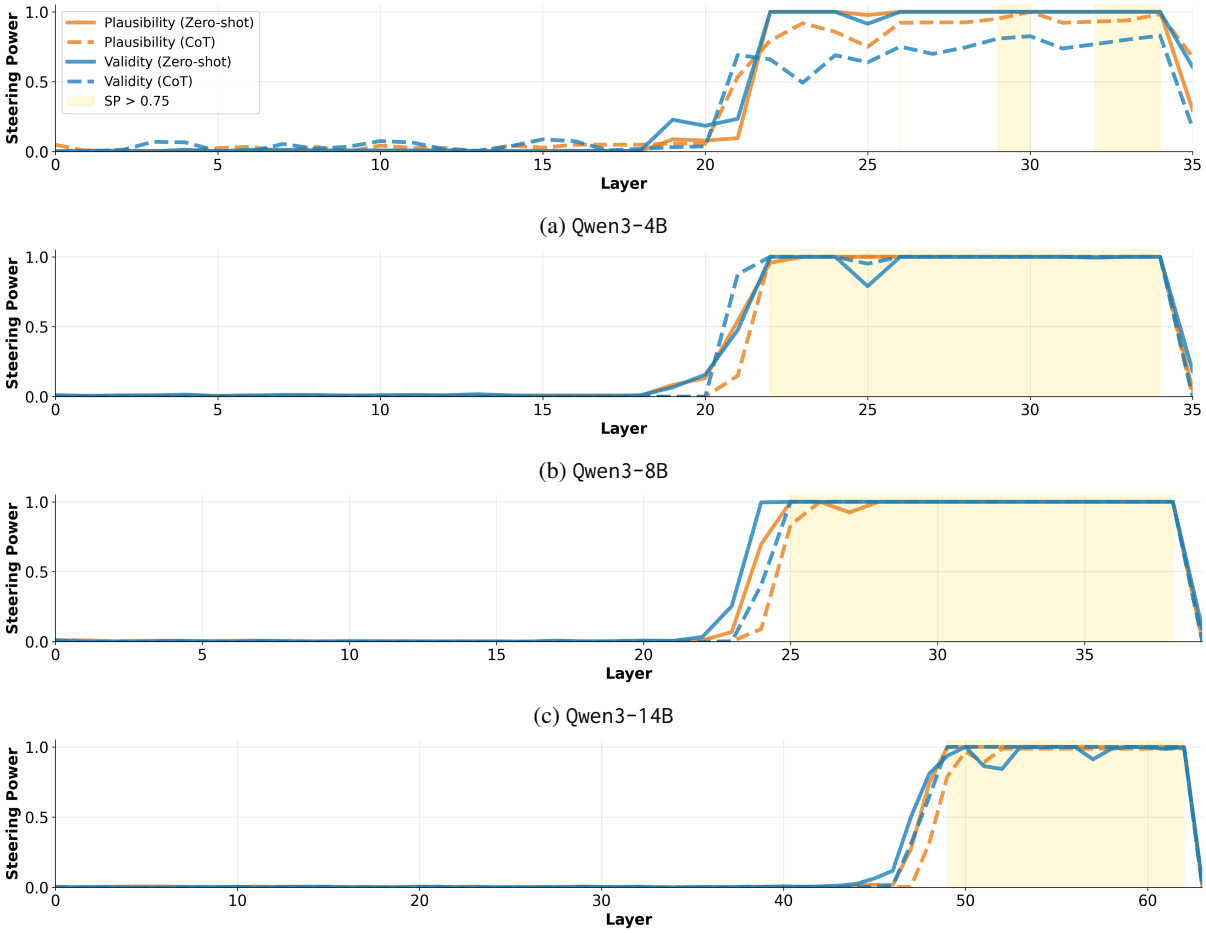


(b) Gemma3-12B-it



(c) Gemma3-27B-it

Figure 11: Steering power (SP) of validity and plausibility vectors applied at different hidden layers across different Gemma-3 model sizes. The yellow regions highlight layers with SP > 0.75 for both validity and plausibility across prompt settings. Validity and plausibility steering vectors exhibit high SP at similar layers under both zero-shot and CoT prompting.



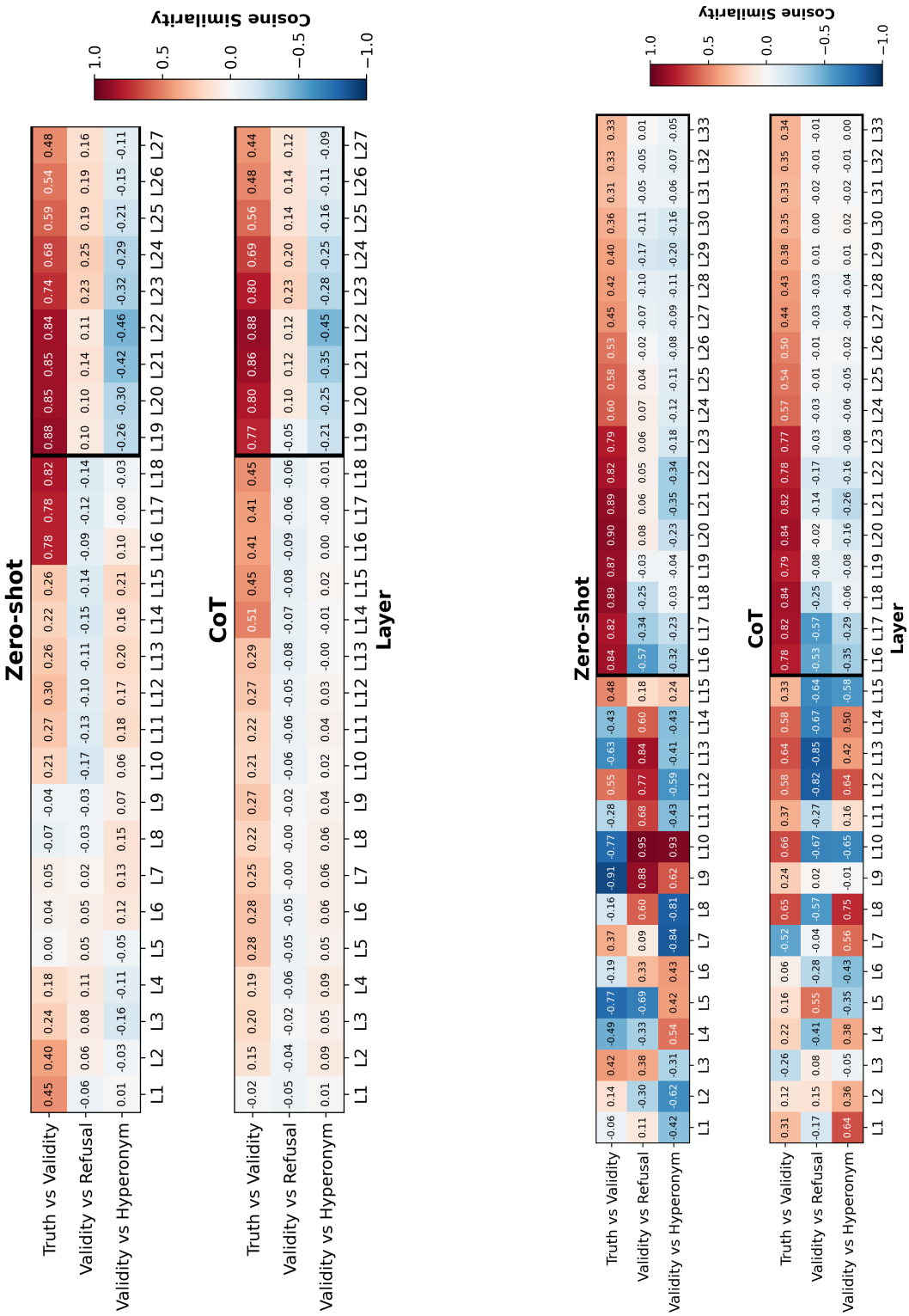
(a) Qwen3-4B

(b) Qwen3-8B

(c) Qwen3-14B

(d) Qwen3-32B

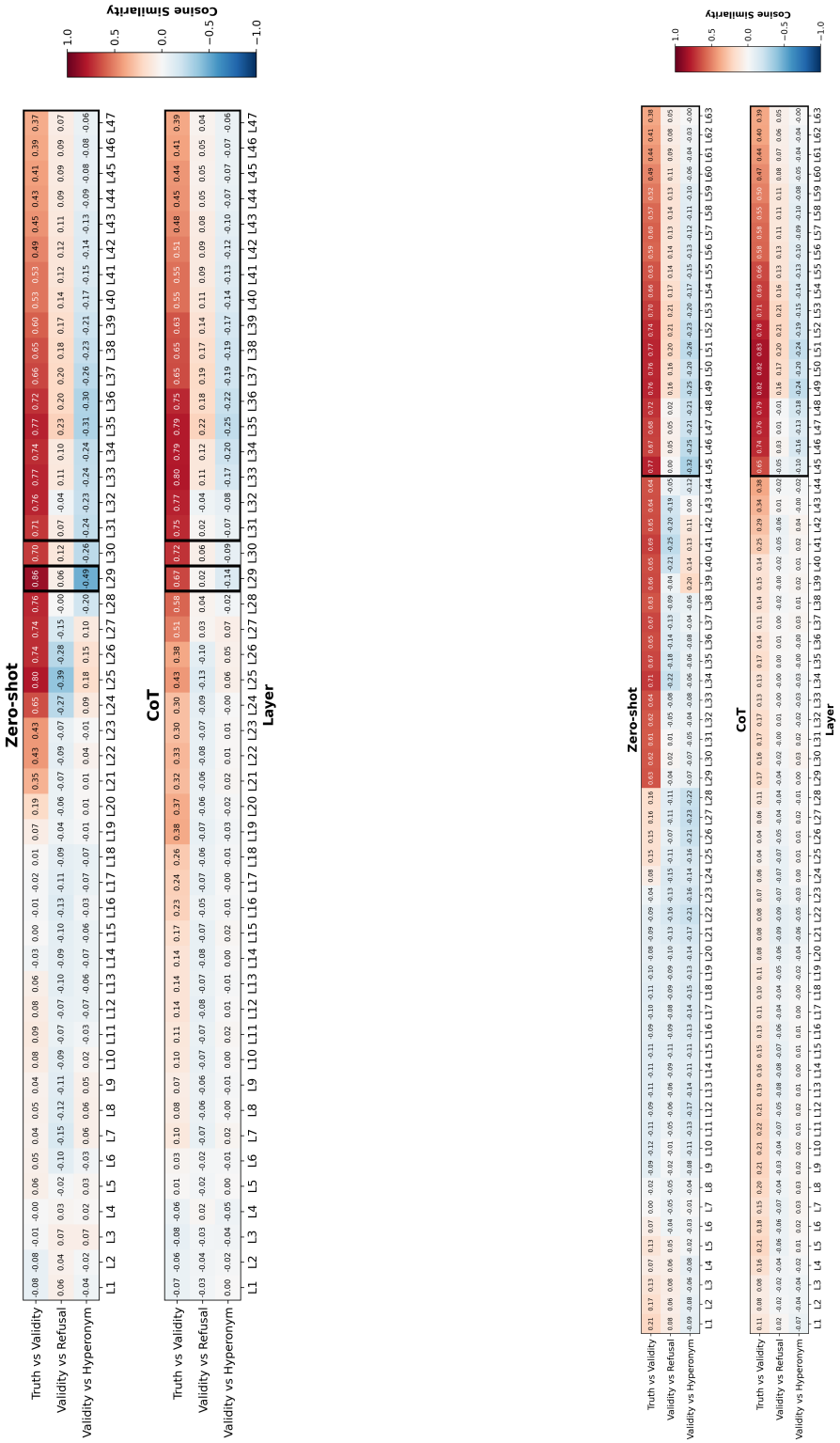
Figure 12: Steering power (SP) of validity and plausibility vectors applied at different hidden layers across different Qwen-3 model sizes. The yellow regions highlight layers with SP > 0.75 for both validity and plausibility across prompt settings. Validity and plausibility steering vectors exhibit high SP at similar layers under both zero-shot and CoT prompting.



(a) Qwen2.5-7B-Instruct

(b) Gemma3-4B-it

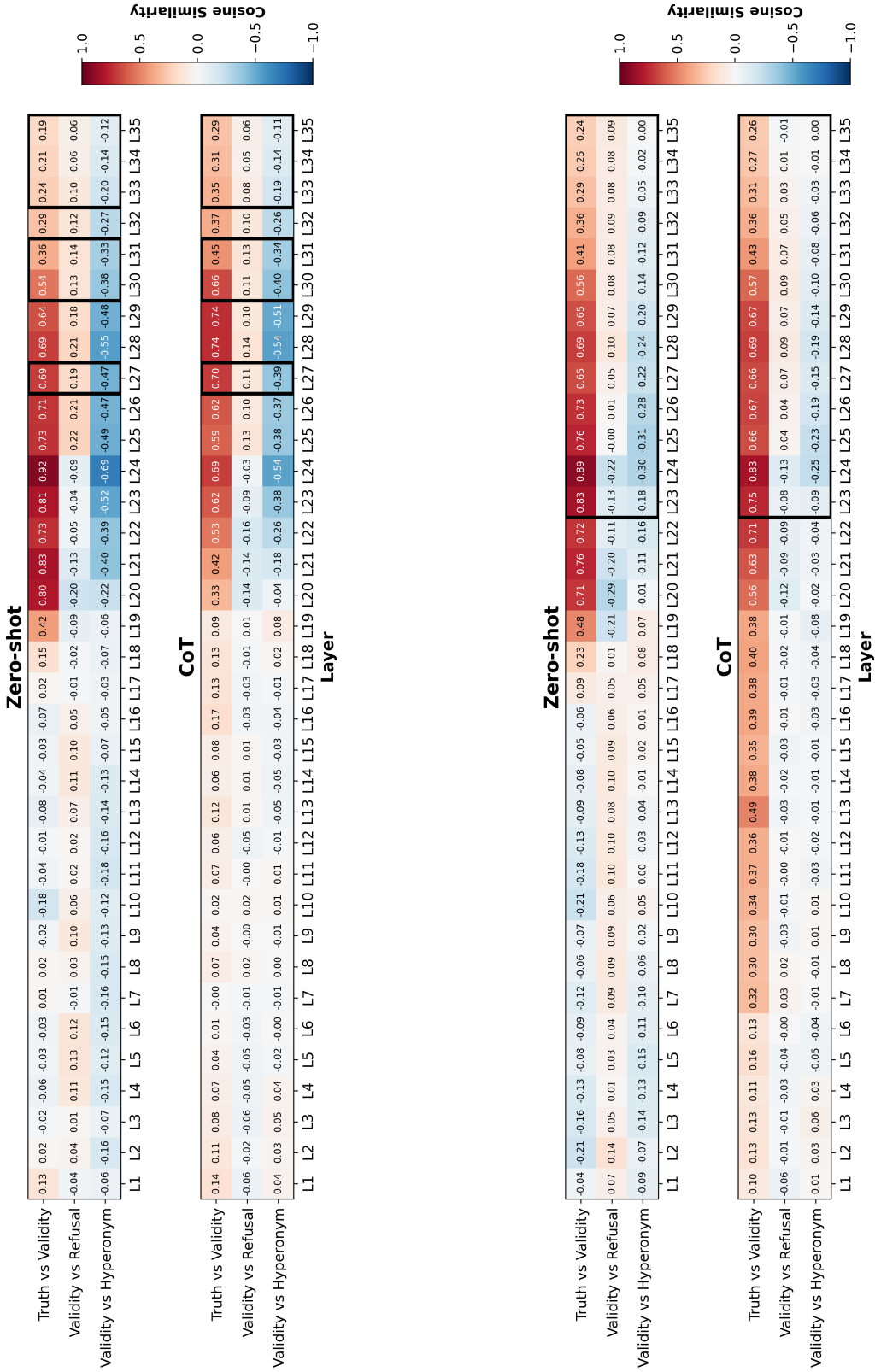
Figure 13: Cosine similarities of vectors representing different concepts with validity vectors extracted at different hidden layers of Qwen2.5-7B-Instruct and Gemma3-4B-it. We consider vectors representing plausibility, harmlessness, and hypernymity. The layers with SP > 0.75 for both validity and plausibility are highlighted using thicker black rectangles. At highly steerable layers, only validity and plausibility have high similarity.



(a) Qwen2.5-14B-Instruct

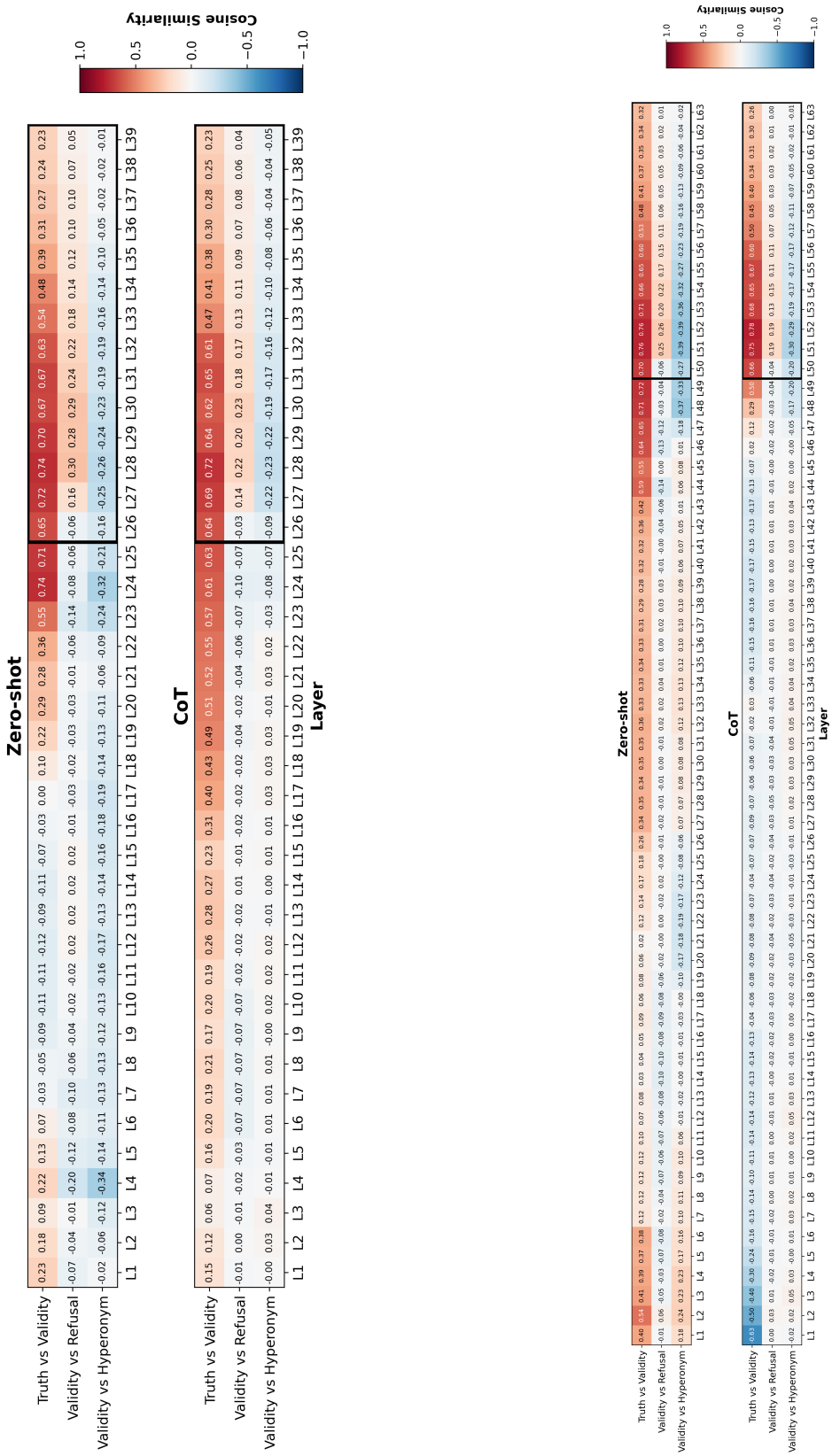
(b) Qwen2.5-32B-Instruct

Figure 14: Cosine similarities of vectors representing different concepts with validity vectors extracted at different hidden layers of Qwen2.5-32B-Instruct and Qwen2.5-14B-Instruct. We consider vectors representing plausibility, harmlessness, and hypernymity. The layers with SP > 0.75 for both validity and plausibility are highlighted using thicker black rectangles. At highly steerable layers, only validity and plausibility have high similarity.



(a) Qwen3-4B (b) Qwen3-8B

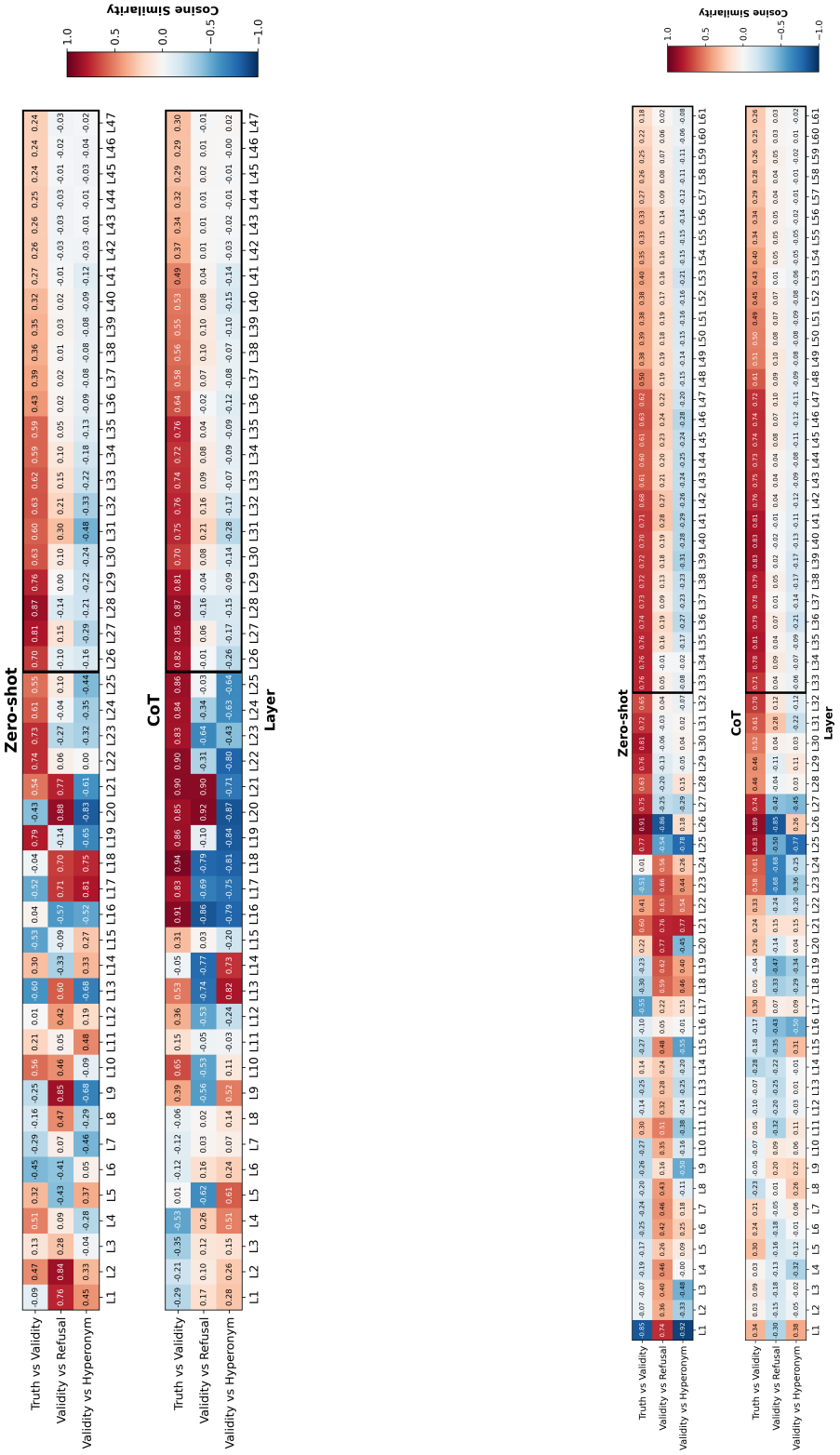
Figure 15: Cosine similarities of vectors representing different concepts with validity vectors extracted at different hidden layers of Qwen3-4B and Qwen3-8B. We consider vectors representing plausibility, harmlessness, and hypernymity. The layers with SP > 0.75 for both validity and plausibility are highlighted using thicker black rectangles. At highly steerable layers, only validity and plausibility have high similarity.



(a) Qwen3-14B

(b) Qwen3-32B

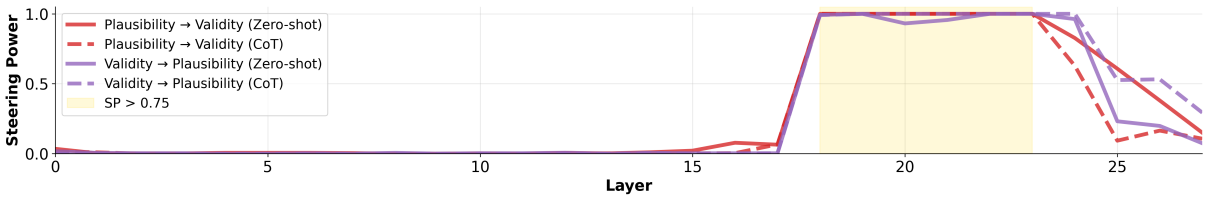
Figure 16: Cosine similarities of vectors representing different concepts with validity vectors extracted at different hidden layers of Qwen3-14B and Qwen3-32B. We consider vectors representing plausibility, harmlessness, and hypernymity. The layers with SP > 0.75 for both validity and plausibility are highlighted using thicker black rectangles. At highly steerable layers, only validity and plausibility have high similarity.



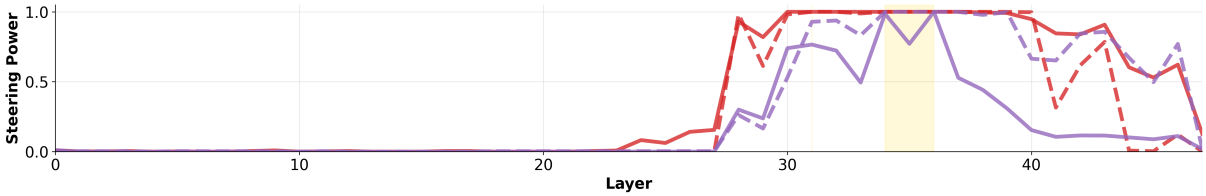
(a) Gemma3-12B-it

(b) Gemma3-27B-it

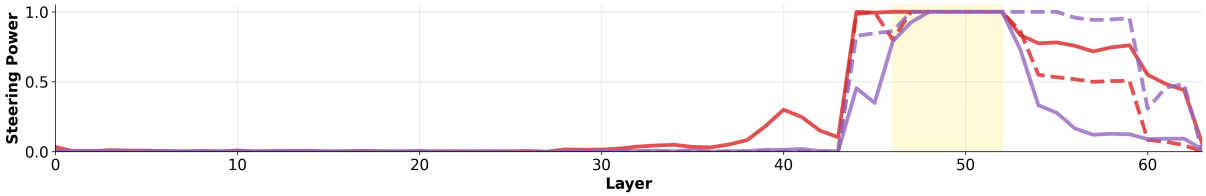
Figure 17: Cosine similarities of vectors representing different concepts with validity vectors extracted at different hidden layers of Gemma3-12B-it and Gemma3-27B-it. We consider vectors representing plausibility, harmlessness, and hypernymity. The layers with SP > 0.75 for both validity and plausibility are highlighted using thicker black rectangles. At highly steerable layers, only validity and plausibility have high similarity.



(a) Qwen2.5-7B-Instruct

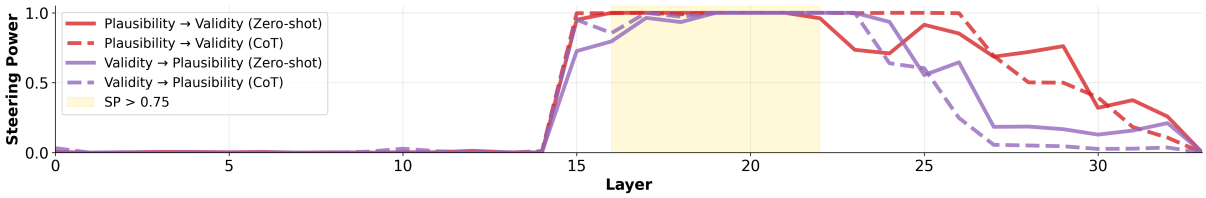


(b) Qwen2.5-14B-Instruct

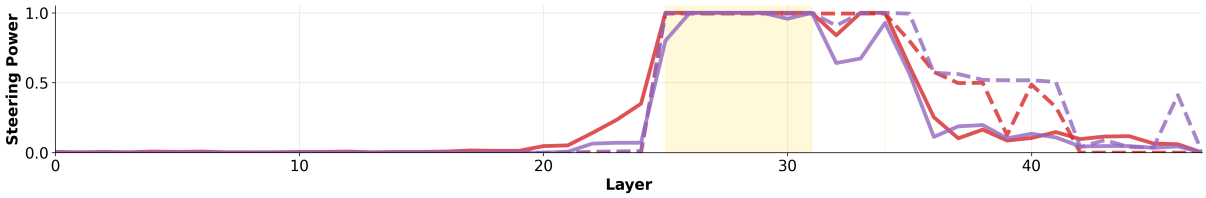


(c) Qwen2.5-32B-Instruct

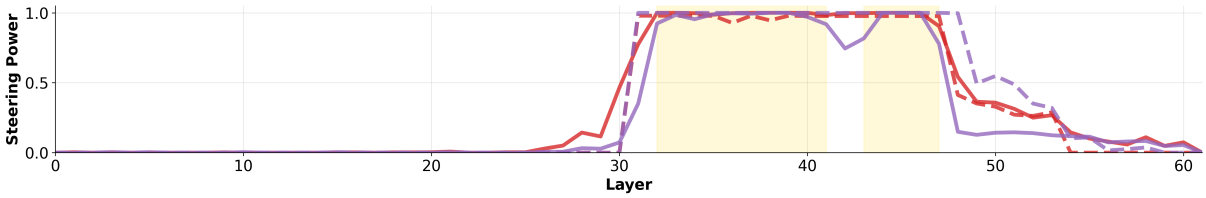
Figure 18: Steering power (SP) of plausibility steering vectors applied to validity classification ("plausibility → validity"), and vice versa ("validity → plausibility") across different hidden layers and Qwen-2.5 model sizes. The yellow regions highlight layers with high transferability in both directions, where SP > 0.75 across prompt settings. High cross-task SP is observed at similar layers under both zero-shot and CoT prompting.



(a) Gemma3-4B-it

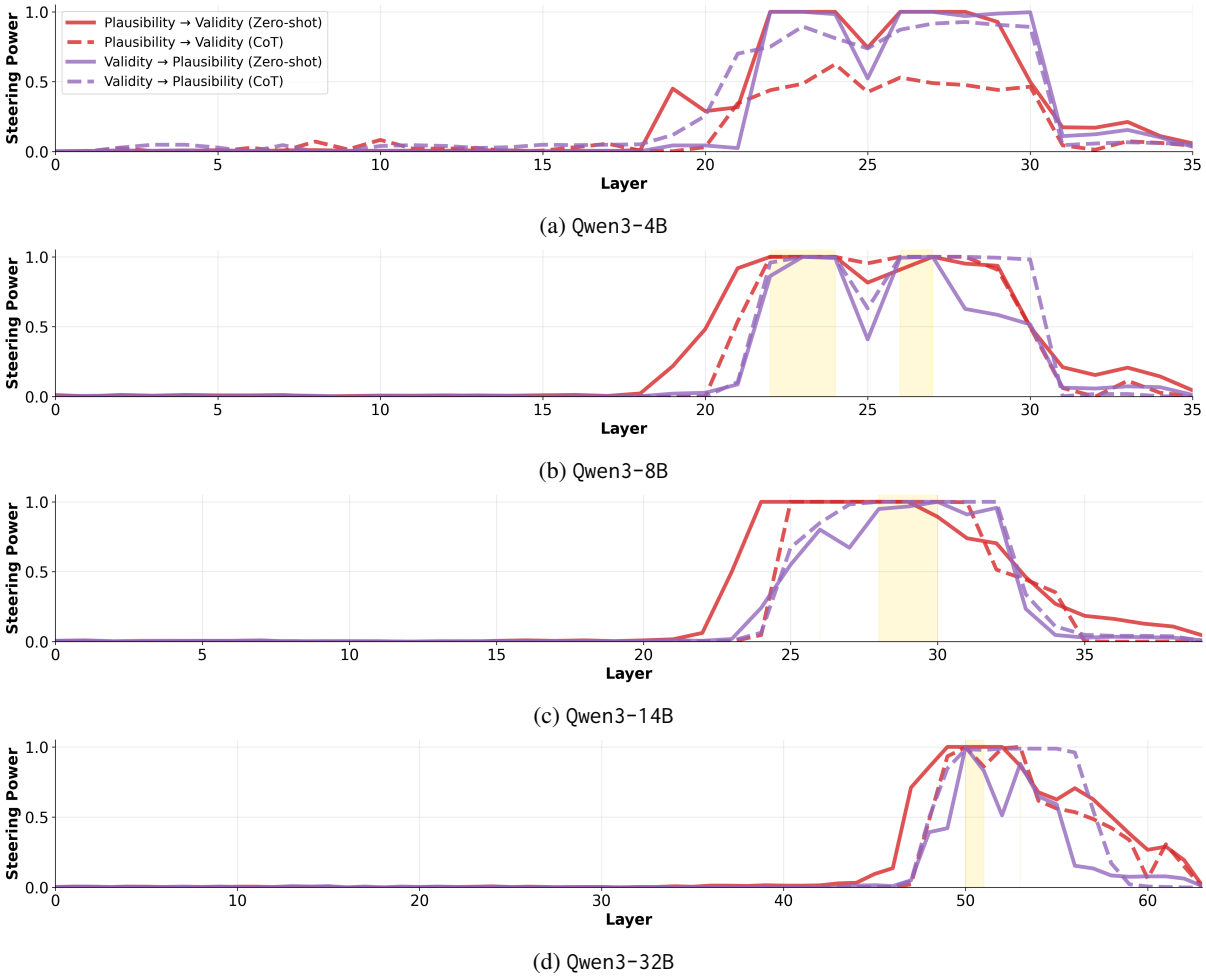


(b) Gemma3-12B-it



(c) Gemma3-27B-it

Figure 19: Steering power (SP) of plausibility steering vectors applied to validity classification ("plausibility → validity"), and vice versa ("validity → plausibility") across different hidden layers and Gemma-3 model sizes. The yellow regions highlight layers with high transferability in both directions, where SP > 0.75 across prompt settings. High cross-task SP is observed at similar layers under both zero-shot and CoT prompting.



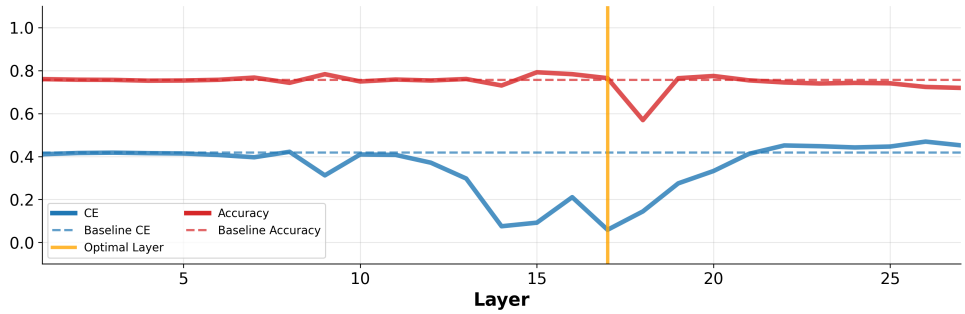
(a) Qwen3-4B

(b) Qwen3-8B

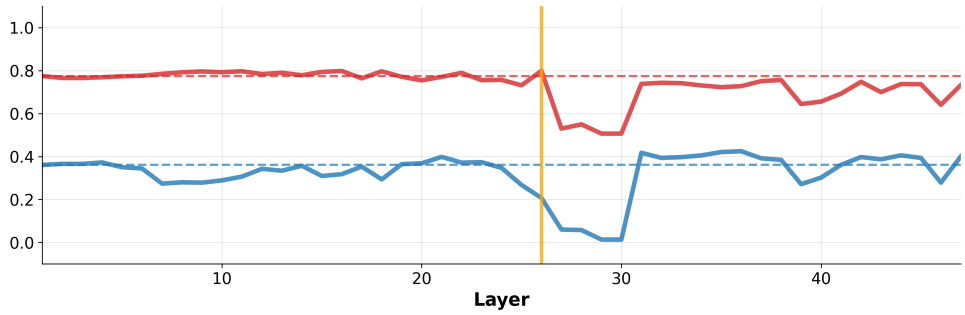
(c) Qwen3-14B

(d) Qwen3-32B

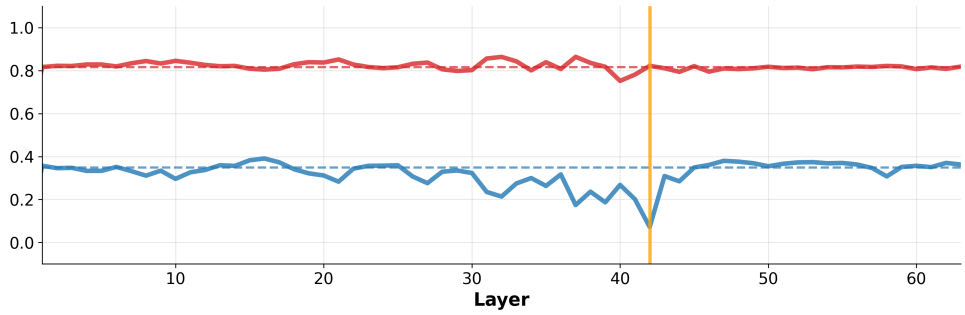
Figure 20: Steering power (SP) of plausibility steering vectors applied to validity classification ("plausibility $\rightarrow$ validity"), and vice versa ("validity $\rightarrow$ plausibility") across different hidden layers and Qwen-3 model sizes. The yellow regions highlight layers with high transferability in both directions, where $\mathrm{SP} > 0.75$ across prompt settings. High cross-task SP is observed at similar layers under both zero-shot and CoT prompting.



(a) Qwen2.5-7B-Instruct

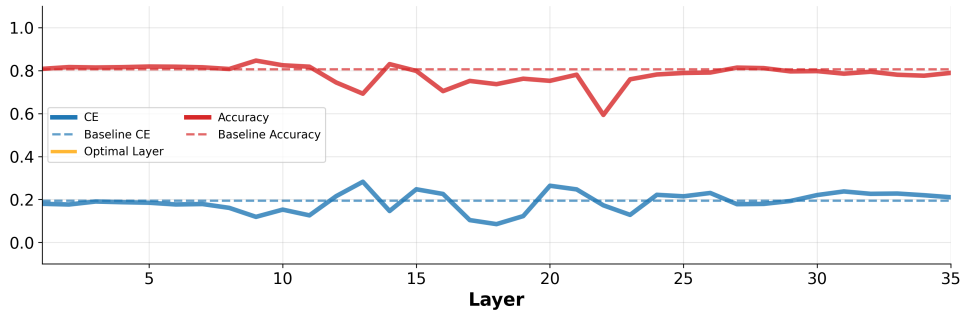


(b) Qwen2.5-14B-Instruct

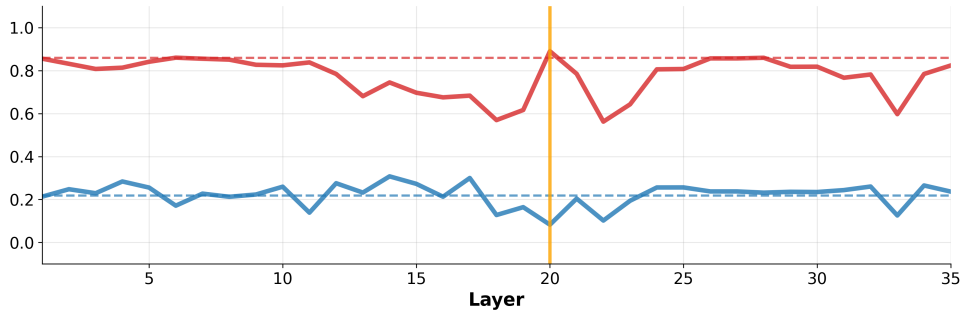


(c) Qwen2.5-32B-Instruct

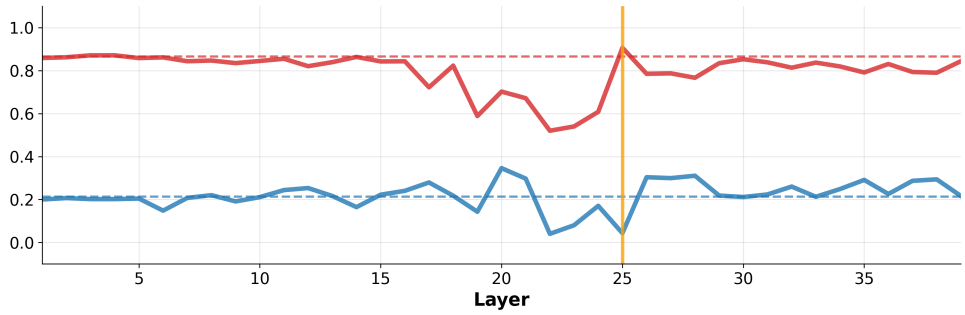
Figure 21: Per-layer accuracy (red), content effect (blue), and success rate (green) of zero-shot Qwen-2.5 models on the logical validity classification task after adding the task difference steering vector μ^l_{V-P} multiplied by a scalar value $\alpha = 1.5$ at different layers. For comparison, original accuracy and content effect are shown as dashed lines. A success rate lower than 1.0 indicates that after the intervention the model did not produce a label for logical validity classification, i.e., $\hat{y} \notin \{\text{valid}, \text{invalid}\}$. The orange line indicates the layer that best retains or improves the original accuracy, while lowering content effect and preserving success rate.



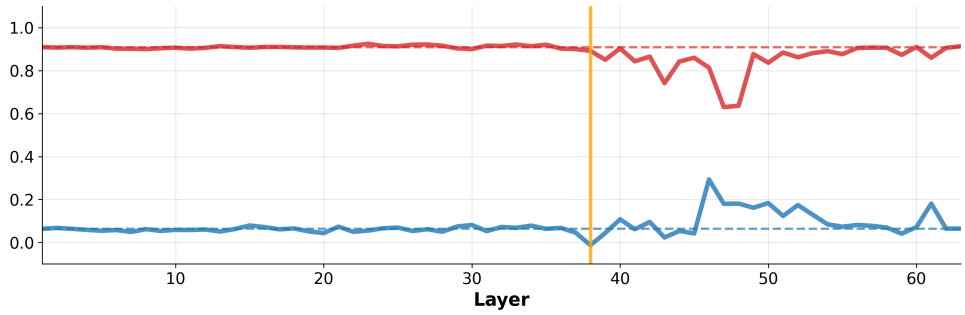
(a) Qwen3-4B



(b) Qwen3-8B

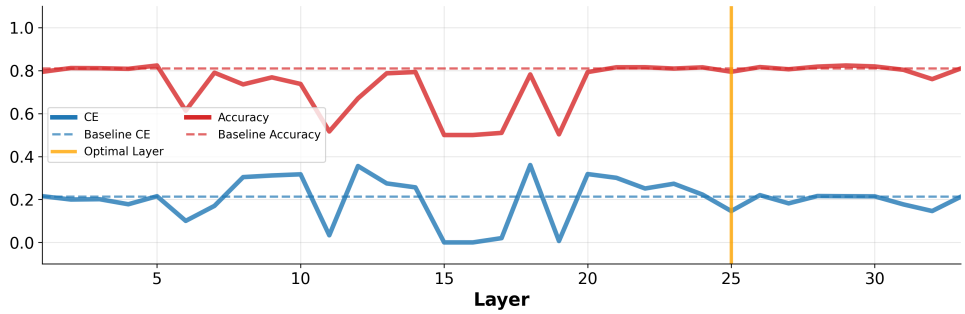


(c) Qwen3-14B

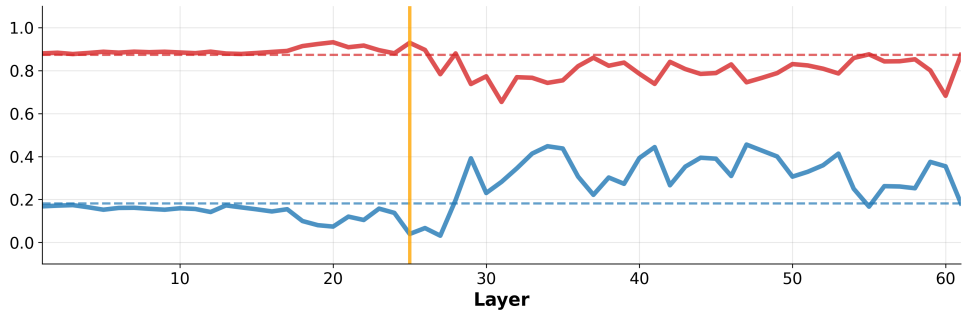


(d) Qwen3-32B

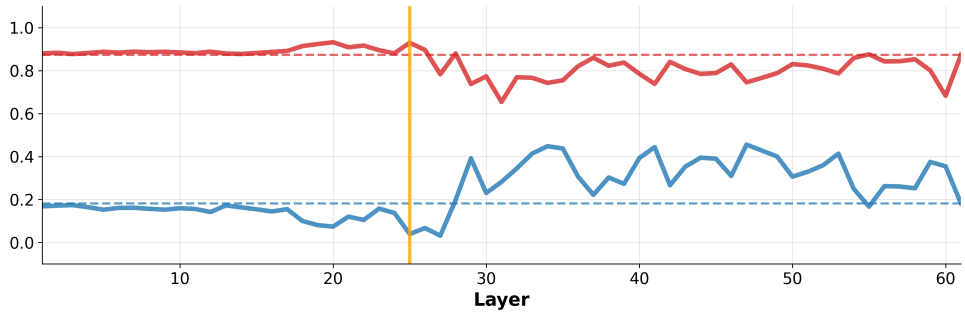
Figure 22: Per-layer accuracy (red), content effect (blue), and success rate (green) of zero-shot Qwen-3 models on the logical validity classification task after adding the task difference steering vector μ^l_{V-P} multiplied by a scalar value $\alpha = 1.5$ at different layers. For comparison, original accuracy and content effect are shown as dashed lines. A success rate lower than 1.0 indicates that after the intervention the model did not produce a label for logical validity classification, i.e., $\hat{y} \notin \{\text{valid}, \text{invalid}\}$. The orange line indicates the layer that best retains or improves the original accuracy, while lowering content effect and preserving success rate.



(a) Gemma3-4B-it



(b) Gemma3-12B-it



(c) Gemma3-27B-it

Figure 23: Per-layer accuracy (red), content effect (blue), and success rate (green) of zero-shot Gemma-3 models on the logical validity classification task after adding the task difference steering vector μ^l_{V-P} multiplied by a scalar value $\alpha = 1.5$ at different layers. For comparison, original accuracy and content effect are shown as dashed lines. A success rate lower than 1.0 indicates that after the intervention the model did not produce a label for logical validity classification, i.e., $\hat{y} \notin \{\text{valid}, \text{invalid}\}$. The orange line indicates the layer that best retains or improves the original accuracy, while lowering content effect and preserving success rate.